COMPLEX NETWORKS 2017

The 6th International Conference on Complex Networks and Their Applications

November 29 - December 01
Lyon, France

BOOK OF ABSTRACTS

COMPLEX NETWORKS 2017
The 6th International Conference on Complex Networks & Their Applications
November 29, 2016 – December 01, 2017
Lyon, France

Published by the International Conference on Complex Networks & Their Applications.

Editors:
Hocine Cherifi
University of Burgundy
France

Hamamache Kheddouci
University of Lyon 1
France

Huijuan Wang
Delft University of Technology
Netherlands

Editorial co-ordination:
Sabrina Gaito
University of Milan
Italy

COMPLEX NETWORKS 2017
e-mail: hocine.cherifi@u-bourgogne.fr

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use. While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained her.

ISBN 978-2-9557050-2-5

Preface

We are proud to present the Book of Abstracts for the 6th International Conference on Complex Networks & their Applications: COMPLEX NETWORKS 2017. Since 2012 the event has been held around the world on a yearly Basis. After Sorrento (Italy), Kyoto (Japan), Marrakech (Morocco), Bangkok (Thailand), Milan (Italy) the sixth edition is hosted by the University of Lyon 2 from November 29 to December 01, 2017. The originality of the conference lies in the strongly interdisciplinary nature of the topics covered. Indeed, complexity and network science are multidisciplinary fields that mobilize intellectual resources in virtually all-scientific communities. Nowadays, all disciplines (physics, biology, social sciences, economics, computer science, meteorology, etc.) are faced with a massive influx of data and an explosion of information to manage. Through the data and their interactions, network science aims at understanding these complex systems increasingly large. COMPLEX NETWORKS is very focused at being an interdisciplinary event. However, this is linked with willingness to the requirements that the quality of the contributions must be among the best work in each of the scientific fields covered. In order to guarantee the excellence and reputation of this event, for its sixth edition COMPLEX NETWORKS has brought together in its scientific committee more than 290 leading international experts from all over the world. Year after year the event has increased its international influence. The 345 contributions that we received this year, from more than 60 countries around the world have been peer reviewed by at least 3 independent reviewers. This publication gathers the 122 extended abstracts accepted for presentation together with abstracts of six keynote speeches and two invited tutorials.

Each edition of the conference represents a challenge that cannot be successfully achieved without the deep involvement of plenty of people, institutions and sponsors. We would like to thank all of them. We record our thanks to our fellow members of the Organizing committee for their huge efforts for the success of the conference. The program committee members for their engagement in promoting the event and refereeing submissions as well as the local committee members for their great commitment over the past months. We are also indebted to our sponsors, in particular Tribe Communication for designing the visual identity of the Conference. We are equally grateful to all the institutions that have helped us, in particular, the University of Lyon 2 for hosting this event. We also wish to express our appreciation to all participants and presenters. On a final note, we would like to express our deep sense of appreciation to our keynote and tutorial speakers.

Lyon,
November 2017

Hocine Cherifi
Sabrina Gaito
Hamamache Kheddouci
Huijuan Wang

Table of Contents

Tutorials

Invited Speakers

I Social Reputation and Influence

II Biological Networks

III Brain Networks

IV Diffusion and Epidemics

V Dynamics on/of Networks

VI Network Models

VII Networks in Finance and Economics

VIII Motif Discovery and Link Analysis

IX Network Analysis and Measures

X Community Structure

XI Resilience and Control

XII Social and Political Networks

Tutorials

Network theory: the challenges that lie ahead

Ginestra Bianconi

Queen Mary University of London

Network theory has emerged almost twenty years ago, as a new field for characterizing interacting complex systems, such as the Internet, the biological networks of the cell, and social networks. This tutorial will provide a (personal) reflection on the maturity of the field, indicating the main results obtained so far and the big challenges that lie ahead. The hot topics that will be critically discussed include: multilayer networks, network geometry and percolation theory.

Ginestra Bianconi is Associate Professor (Reader) and Director of the MSc in Network Science at the School of Mathematical Sciences at Queen Mary University of London, London, UK. Her research activity on network science includes network theory and its applications and has appeared in journal such as Science, PNAS, PRX and Physical Review Letters. In the last years her work have focused on multilayer networks, network geometry, percolation and network control.

Mining Information Propagation Data

Francesco Bonchi -

ISI Foundation

With the success of online social networks and microblogging platforms such as Facebook, Tumblr, and Twitter, the phenomenon of influence-driven propagations, has recently attracted the interest of computer scientists, sociologists, information technologists, and marketing specialists. In this talk we will take a data mining perspective, discussing what (and how) can be learned from a social network and a database of traces of past propagations over the social network. Starting from one of the key problems in this area, i.e. the identification of influential users, we will provide a brief overview of our recent contributions in this area. We will expose the connection between the phenomenon of information propagation and the existence of communities in social network, and we will go deeper in this new research topic arising at the overlap of information propagation analysis and community detection.

Francesco Bonchi is Research Leader at the ISI Foundation, Turin, Italy, where he's the head of the ”Algorithmic Data Analytics” group. He is also (part-time) Principal Scientist for Data Mining at Eurecat (Technological Center of Catalunya),Barcelona. Before he was Director of Research at Yahoo Labs in Barcelona, Spain, where he was leading the Web Mining Research group.

His recent research interests include mining query-logs, social networks, and social media, as well as the privacy issues related to mining these kinds of sensible data. In the past he has been interested in data mining query languages, constrained pattern mining, mining spatiotemporal and mobility data, and privacy preserving data mining.

He is member of the ECML PKDD Steering Committee, Associate Editor of the newly created IEEE Transactions on Big Data (TBD), of the IEEE Transactions on Knowledge and Data Engineering (TKDE), the ACM Transactions on Intelligent Systems and Technology (TIST), Knowledge and Information Systems (KAIS), and member of the Editorial Board of Data Mining and Knowledge Discovery (DMKD). He has been program co-chair of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010). Dr. Bonchi has also served as program co-chair of the 28th ACM Conference on Hypertext and Hypermedia (HT 2017), the 16th IEEE International Conference on Data Mining (ICDM 2016), the first and second ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD 2007 and 2008), the 1st IEEE International Workshop on Privacy Aspects of Data Mining (PADM 2006), and the 4th International Workshop on Knowledge Discovery in Inductive Databases (KDID 2005). He is co-editor of the book ”Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques” published by Chapman & Hall/CRC Press. He earned his Ph.D. in computer science from the University of Pisa in December 2003.

Invited Speakers

Finding the most versatile nodes in highly multidimensional data

Alex Arenas

Universitat Rovira i Virgili

The determination of the most central agents in complex networks is important because they are responsible for a faster propagation of information, epidemics, failures and congestion, among others. A challenging problem is to identify them in networked systems characterized by different types of interactions, forming interconnected multilayer networks. Here we describe a mathematical framework that allows us to calculate centrality in such networks and rank nodes accordingly, finding the ones that play the most central roles in the cohesion of the whole structure, bridging together different types of relations. These nodes are the most versatile in the multilayer network. We investigate empirical interconnected multilayer networks and show that the approaches based on aggregating—or neglecting—the multilayer structure lead to a wrong identification of the most versatile nodes, overestimating the importance of more marginal agents and demonstrating the power of versatility in predicting their role in propagation processes with applications in social networks, banking networks, etc.

Prof. Alex Arenas (Barcelona, 1969) got his PhD in Physics in 1996. In 1995, he got a tenure position at Dept. Computer Science and Mathematics (DEIM) at Universitat Rovira i Virgili, and in 1997 he became associate professor at the same department. In 2000, he was visiting scholar at the Lawrence Berkeley Lab. (LBL) in the Applied Mathematics group of Prof. Alexandre Chorin (University of California, Berkeley). After this visit, he started a collaboration with Berkeley, and in 2007 he became visiting researcher of LBL. Arenas has written more than 160 interdisciplinary publications in major peer reviewed including Nature, Nature Physics, PNAS, Physics Reports and Physical Review Letters, which have received more than 9000 citations. He is one of the few Europeans serving as Associate Editors of the most important publication in physics worldwide, the American Physical Society journal, Physical Review. He is in charge of the Complex Networks and Interdisciplinary Physics section of Physical Review E. He got the James Mc Donnell Foundation award for the study of complex systems in 2011. He was also recognized as ICREA Academia-Institució Catalana de Recerca i Estudis Avançats, a catalan award that promotes the most recognized scientists from Catalonia. He serve as Editor in Journal of Complex Networks, and in Network Neuroscience. He was elected for the Steering Committee of the Complex Systems Society in 2012. He is the leader of the research group ALEPH-SYS.

Good City Life

Daniele Quercia

Nokia Bell Labs

The corporate smart-city rhetoric is about efficiency, predictability, and security. “You’ll get to work on time; no queue when you go shopping, and you are safe because of CCTV cameras around you”. Well, all these things make a city acceptable, but they don’t make a city great. We are launching goodcitylife.org - a global group of like-minded people who are passionate about building technologies whose focus is not necessarily to create a smart city but to give a good life to city dwellers. The future of the city is, first and foremost, about people, and those people are increasingly networked. We will see how a creative use of network-generated data can tackle hitherto unanswered research questions. Can we rethink existing mapping tools (`http://www.ted.com/talks/daniele_quercia_happy_maps`)? Is it possible to capture smellscapes of entire cities and celebrate good odors (`http://goodcitylife.org/smellymaps/index.html`)? And soundscapes (`http://goodcitylife.org/chattymaps/index.html`)?

Daniele Quercia is a computer scientist and is currently building the Social Dynamics team at Bell Labs Cambridge UK, has been named one of Fortune magazine’s 2014 Data All-Stars, and spoke about “happy maps” at TED. His research area is urban computing. His research received best paper awards from ACM Ubicomp 2014 and from AAAI ICWSM 2015, and an honorable mention from AAAI ICWSM 2013. He was Research Scientist at Yahoo Labs, a Horizon senior researcher at The Computer Laboratory of the University of Cambridge, and Postdoctoral Associate at the Massachusetts Institute of Technology. He received his PhD from UC London. His thesis was sponsored by Microsoft ResearchCambridge and was nominated for BCS Best British PhD dissertation in Computer Science. During his PhD, he was MBA Technology Fellow at London Business School.

Community structure in complex networks

Santo Fortunato

Indiana University

Complex systems typically display a modular structure, as modules are easier to assemble than the individual units of the system, and more resilient to failures. In the network representation of complex systems, modules, or communities, appear as subgraphs whose nodes have an appreciably larger probability to get connected to each other than to other nodes of the network. In this talk I will address three fundamental questions: How is community structure generated? How to detect it? How to test the performance of community detection algorithms? I will show that communities emerge naturally in growing network models favoring triadic closure, a mechanism necessary to implement for the generation of large classes of systems, like e.g. social networks. I will discuss the limits of the most popular class of clustering algorithms, those based on the optimization of a global quality function, like modularity maximization. Testing algorithms is probably the single most important issue of network community detection, as it implicitly involves the concept of community, which is still controversial. I will discuss the importance of using realistic benchmark graphs with built-in community structure, as well as the role of metadata.

He received his PhD in Theoretical Physics in 2000 at the Department of Physics of the University of Bielefeld, Germany, working on lattice gauge theories, percolation and phenomenology of heavy-ion collisions. He switched to complexity science in 2004, and from 2005 till 2007 he has been postdoctoral researcher at the School of Informatics and Computing of Indiana University, working in the group of Alessandro Vespignani. From 2007 till 2011 he has been at ISI Foundation in Turin, Italy, first as research scientist then as a scientific leader. In 2011 he became Associate Professor in Complex Systems at the School of Science of Aalto University , Finland. He is currently full professor in the School of Informatics and Computing at Indiana University.

The impact of network structure on relational machine learning

Jennifer Neville

Purdue University

Network science focuses on analyzing network structure in order to understand key relational patterns in complex systems. In contrast, relational machine learning typically conditions on the observed relations in a network, using them as a form of inductive bias to constrain the space of dependencies considered by the models. While recent interest in these two fields has produced a large body of research on models of both network structure and relational data, there has been less attention on the intersection of the two fields–specifically considering the impact of network structure on relational learning methods. Since many relational domains comprise a single, large, partially-labeled network, many of the conventional assumptions in relational learning are no longer valid and the network structure creates unique statistical challenges for learning and inference algorithms. This talk will outline some of the algorithmic and statistical challenges that arise due to partially-observed, large-scale networks, and describe methods for semi-supervised learning, latent-variable modeling, and sampling to address the challenges.

Jennifer Neville is the Miller Family Chair Associate Professor of Computer Science and Statistics at Purdue University. She received her PhD from the University of Massachusetts Amherst in 2006. She is currently an elected member of the AAAI Executive Council and she was recently PC chair of the 9th ACM International Conference on Web Search and Data. In 2012, she was awarded an NSF Career Award, in 2008 she was chosen by IEEE as one of ”AI’s 10 to watch”, and in 2007 was selected as a member of the DARPA Computer Science Study Group. Her work, which includes more than 100 publications with over 5000 citations, focuses on developing data mining and machine learning techniques for complex relational and network domains, including social, information, and physical networks.

Spreading influence in social networks: From link-centric to node-centric models

Frank Schweitzer

ETH Zürich

Epidemic spreading on complex networks is well studied because nodes follow a rather simple dynamics. Thus, the focus is mostly on how the network topology impacts the spreading process. However, modeling the spread of, e.g., emotions in online social networks requires us to have more refined models of the node dynamics, to calculate cascades of spreading influence. We capture the node dynamics by means of a data-driven modeling approach that allows us to test, and to calibrate, assumptions about the user behavior. In my talk, I present different examples of how to complement the topological perspective by a node-centric perspective that considers costs and benefits, emotional responses or information processing of users.

Frank Schweitzer has been Full Professor for Systems Design at ETH Zurich since 2004. He is also associated member of the Department of Physics at the ETH Zurich. Frank Schweitzer received his first Ph.D. (Dr. rer. nat.) in theoretical physics at the age of 26, and his second Ph.D. (Dr. phil.) in philosophy of science at the age of 29, he further earned a habilitation/Venia Legendi in Physics. In his professional career, he worked for different research institutions (Max-Planck Institute for the Physics of Complex Systems, Dresden, Fraunhofer Institute for Autonomous Intelligent Systems, Sankt Augustin) and universities (Humboldt University Berlin, Cornell University Ithaca NY, Emory University, Atlanta GA).

The research of Frank Schweitzer focuses on applications of complex systems theory in the dynamics of social and economic organizations. He is interested in phenomena as diverse as user interaction in online social networks, collective decisions in animal groups, failure cascades and systemic risk in economic networks, and the rise and fall of collaborations in socio-technical systems. His methodological approach can be best described as data-driven modeling, i.e., it combines the insights from big data analysis with the power of agent-based computer simulations and the strength of rigorous mathematical models. Frank Schweitzer is a founding member of the ETH Risk Center and Editor-in-Chief of ACS - Advances in Complex Systems and EPJ Data Science.

Network analysis literacy - a socioinformatic approach

Katharina A. Zweig

TU Kaiserslautern

Why are there so many centrality indices? This is the question that puzzled me when I started into network analysis in 2003. Borgatti showed that centrality indices are best understood as tightly coupled to a specific kind of network flow or network process associated with it. His main idea, that centrality indices come with a model of a network flow or process, can be generalized to other types of data mining and quality measures. I will thus discuss the question of responsibility when measures are used in societally important algorithmic decision making systems, such as terrorist identification systems which include social network features.

Katharin A. Zweig is a professor at the TU Kaiserslautern since 2012. As a studied biochemist and computer scientist, her postdoc was in the biophysics group of Prof. Dr. Tamás Vicsek at ELTE University Hungary. With this interdisciplinary background, she designed a new field of study called Socioinformatics at the TU Kaiserslautern. It is concerned with the impact of IT Systems on individuals, organizations, and society at large. In her research, Zweig first focused on understanding when to use which network analytic measure for a meaningful interpretation of the result. Her research has now broadened to the meaningful use of other types of data mining. She is a junior fellow of the German Society of Computer Science from 2013 (until 2018), was selected as a ”Digital Head” in 2014 in Germany, and won the ars-legendi teaching prize in Engineering and Computer Science in 2017. She co-founded an initiative called ”Algorithm Watch” in 2016 and counsels politics, churches, media authorities and foundations with respect to the impact of algorithms on society.

Part I

Social Reputation and Influence

Hierarchical and Circulating Flow Structure in an Interfirm Transaction Network

Yuichi Kichikawa[1], Hiroshi Iyetomi[1], Takashi Iino[1], and Hiroyasu Inoue[2]

[1]Faculty of Science, Niigata University, Niigata 950-2181, JAPAN
[2]Graduate School of Simulation Studies, University of Hyogo, Kobe 650-0047, JAPAN

1 Introduction

In general, interactions between individuals are considered to play an important role in the economy. For instance, firms are connected to each other directly or indirectly through their business transactions. A firm obtains materials from suppliers and sells its products to customers. These transactions are so essential to firms that one cannot isolate the dynamics of individual firms from the entire economic system.

Conventionally, the industrial structure and economic ripple effects have been studied on the basis of the input-output tables [1]. Furthermore, a network-theoretic point of view was incorporated into the input-output analysis to elucidate complex inter-industrial flow structures [2, 3]. However, such classification of firms by industry may be too formal for a reliable macroscopic picture of the economy.

The objective of this study is to shed an empirical light on industrial flow structures embedded in microscopic supplier-buyer relations. We first construct a directed network from actual data of interfirm transaction relations. And then we analyze the flow structure of the network with a special emphasis on its hierarchy and circularity.

2 Data and Method

The present analysis is based on a big data of 4,974,802 transaction relations between 1,066,037 firms in Japan, collected by the Tokyo Shoko Research, Ltd. in 2016.[1] This data virtually covers whole industrial activities in Japan. We regard firms as nodes and transaction relations between them as directed links spanning from suppliers to customers. Also, we assume all the links have the same weight; information on the volume of each transaction is not available.

The directed network thus constructed is significantly heterogeneous. To extract important structural information in the network, we utilize the map equation [4]. It detects such communities within which nodes are strongly tied bidirectionally and hence exposed to common shocks with large probability; in return, the direction of flow across the communities is biased in an either way. Also we apply the Helmholtz-Hodge decomposition [5] to the transaction network with firms or communities as nodes, decomposing flow on a network into two components: potential flow and loop flow. The potential of nodes identifies their hierarchical positions in the flow structure. The loop flow component may illuminate circulating feedback built in the industrial system.

[1] This is the largest connected component in the network obtained from the original data, containing 99.3% of all active firms listed in the data.

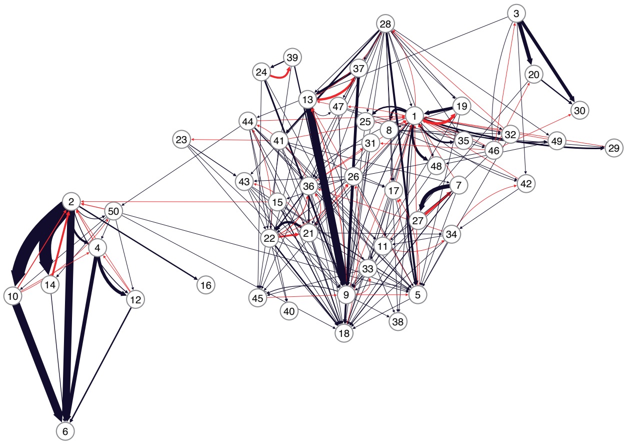


Fig. 1. Visualization of the interfirm transaction network in which the communities are represented simply as nodes. Here only the 50 largest communities are displayed. Node numbers are sorted in descending order of community size. The width of each directed link for a pair of nodes reflects the number of transactions across the pair. The nodes are aligned vertically with their potential values obtained by the Helmholtz-Hodge decomposition for the coarse-grained network. The blue arrows depict flow from upstream to downstream nodes and the red arrows, vice versa.

3 Results and Discussion

The network under study is decomposable into 80,092 communities by the map equation. A vast range of looping flow components are identified as communities, and major communities are then characterized in terms of their regional, industry, and business affiliations. In contrast, the transactions are sparse or polarized across the communities. We construct a coarse-grained network with the communities as nodes and calculate the potential of each community by the Helmholtz-Hodge decomposition.

Figure 1 is a diagram of the communities which are arranged vertically with their potential values; basically goods and services flow from top to bottom on the diagram. From this figure, one can sketch the hierarchical and circulating flow structure in the interfirm transaction network. For instance, nodes 2, 6, 10, and 14 form a subgroup, all dominated by medical, health care & welfare firms, but there is significant hierarchy among them. Also nodes 3, 20, and 30, dominated by whole sales & retail trade, form a rather isolated cluster. In the main body of the network, there is a large flow from node 13 to node 9. The former contains firms of whole sales & retail trade, construction, and manufacturing, big three industries in Japan, at the average composition. The latter is dominated by construction.

Figure 2 shows distributions of the potential values for firms classified by major category of industry. The average potential value of each distribution describes hierarchical characteristics of the corresponding industry. The manufacturing industry is located at the upstream side as compared with the service industry, etc. This is consistent with the general idea on the supply chain. However, even within the same industry, potentials are widely distributed from upstream to downstream except for finance & insurance, medical, health care & welfare, and government.

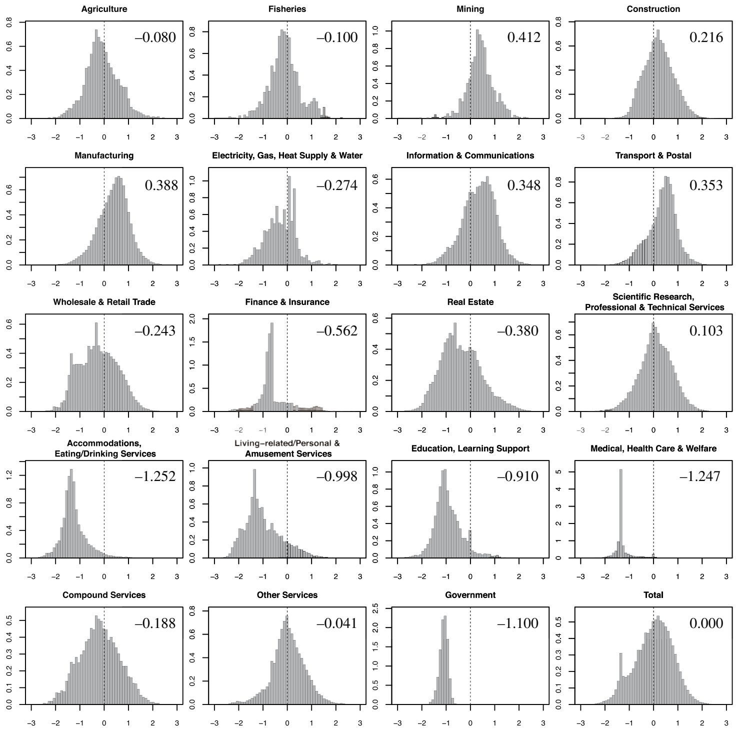


Fig. 2. Distributions of the Helmholtz-Hodge potential for firms in the industry divisions. Firms with high potential values are positioned at the upstream side in the supply chain. The figure in each panel is the averaged value of the potential over its distribution.

These results suggest limitations of the conventional industrial classification scheme for firms in analyzing economic activities, which may be replaced by a new one based on the flow-based community structure in the actual interfirm transaction network.

This research was supported by MEXT as Exploratory Challenges on Post-K computer (Studies of Multi-level Spatiotemporal Simulation of Socioeconomic Phenomena), by a grant-in-aid from Zengin Foundation for Studies on Economics and Finance, and by JSPS KAKENHI Grant Number 16K13375.

References

1. Leontief, Wassily W.: Input-Output Economics. Oxford University Press (1986)
2. Slater, P.: The determination of groups of functionally integrated industries in the United States using a 1967 interindustry flow table. Empirical Economics 2.1, 1-9 (1977); The network structure of the United States input-output table. ibid. 3.1, 49-70 (1978)
3. McNerney, James, Brian D. Fath, and Gerald Silverberg.: Network structure of inter-industry flows. Physica A: Statistical Mechanics and its Applications 392.24, 6427-6441 (2013)
4. Rosvall, Martin, and Carl T. Bergstrom.: Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PLoS ONE 6.4, e18209 (2011)
5. Jiang, Xiaoye, et al.: Statistical ranking and combinatorial Hodge theory. Mathematical Programming 127.1, 203-244 (2011)

Excess reciprocity distorts reputation in online social networks

Giacomo Livan[1,2], Fabio Caccioli[1,2], and Tomaso Aste[1,2]

[1] Department of Computer Science, University College London, 66-72 Gower Street, WC1E 6EA London (UK)
[2] Systemic Risk Centre, London School of Economics and Political Sciences, Houghton Street, WC2A 2AE London (UK)

1 Introduction

Peer-to-peer (P2P) platforms rely on trust between users who have never interacted before neither offline nor online. Trust is typically ensured by requiring users to build a reputation score through digital peer-review mechanisms that allow users to rate their peers. Given the increasingly prominent role the P2P paradigm is playing in the digital economy, online reputation will become more and more central in our online lives. So, it is crucial to ensure that digital peer-review systems produce reliable reputation scores.

Their current lack of regulation exposes P2P platforms to a number of biases. Game theoretic considerations [1], and plenty of anecdotal evidence, suggest that users are often incentivized to reciprocate ratings in order to boost or damage each other's reputations. For instance, the "5 for 5" practice of Uber drivers and passengers, i.e. agreeing on exchanging 5 star ratings at the beginning of a ride, is a common firsthand experience of such practices, and similar phenomena have been reported in the interactions between eBay buyers and sellers [2] and Airbnb hosts and guests [3].

We investigate this issue in three P2P platforms where ratings are binary (i.e. a rating corresponds to either a "like" or a "dislike"), which can be conveniently represented in terms of signed networks: Slashdot, a platform where users comment technology related news and can label each other as "friends" or "foes", Epinions, a platform for crowdsourced product reviews whose users rate the usefulness of each other's reviews, and a collection of interpreted interactions on a set of Wikipedia pages. We investigate the impact of reciprocity on the aggregate scores that representing users' reputations, and assess how their ranking is affected by reciprocity. Our results have been published in [4], and we summarize them in the following.

2 Results

1. In all three platforms we detect excess reciprocity, both in positive and negative ratings, with respect to ensembles of null network models obtained through a link rewiring procedure that partially randomizes the original networks while retaining *i*) the overall heterogeneity in the networks, *ii*) the reputation of each individual user (quantified by the normalized difference between the number of incoming positive and negative links

/ ratings), and *iii*) a proxy of user homophily which takes into account the natural tendency of users to like / dislike the same peers.

2. We complement the above finding by testing whether the networks could sustain higher levels of reciprocity. We find that positive reciprocity is almost “at capacity”, i.e. the networks could not accommodate much more reciprocity than they already display. We interpret this finding as evidence that the growth of P2P platforms is dominated by the exchange of reciprocated ratings.

3. We find that the contribution to reputation from reciprocated (unreciprocated) ratings in empirical networks is systematically over-expressed (under-expressed) with respect to all the null hypotheses we investigate (see Fig. 1). We refer to this as *reciprocity bias*, and we propose a simple link removal algorithm as a benchmark for a possible platform management policy to counteract it.

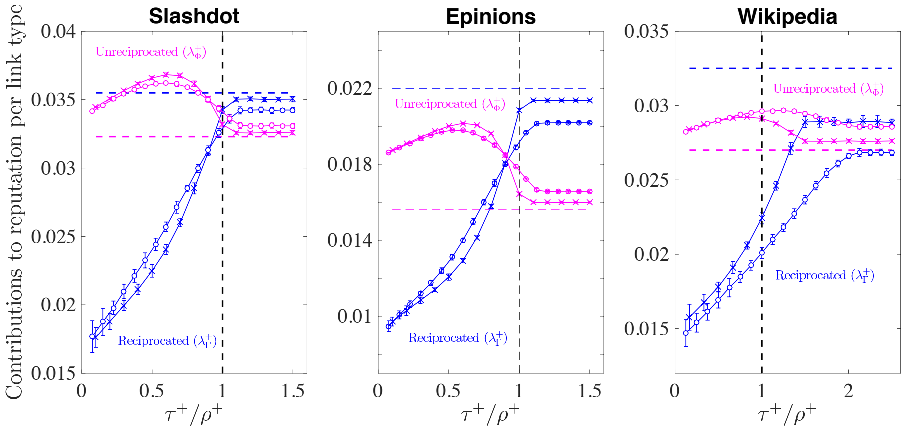


Fig. 1. Demonstration that reputation is affected by reciprocity bias. In each plot blue (pink) solid lines show the average contribution to user reputation from one reciprocated (unreciprocated) rating in a null model ensemble. Blue (pink) dashes lines show the corresponding values measured in the empirical networks. Values are shown as a function of the parameter τ^+, which tunes reciprocity in the null models, divided by the reciprocity value ρ^+ measured in the empirical networks.

Summary. Our results shed light on the processes underpinning the formation of online reputation in decentralized online environments. Being of a purely statistical nature, they only allow us to make claims about the likelihood of detecting the patterns that are indeed empirically observed, not on the social and psychological dynamics leading to their emergence. The latter can be instead investigate through behavioral experiments in controlled environments, which are currently being run by the authors and will be the subject of a future publication.

References

1. Bolton, G., Greiner, B., Ockenfels, A.: Engineering trust: reciprocity in the production of reputation information, Manage. Sci. 59, 265-285 (2013)
2. Resnick, P., Zeckhauser, R., Swanson, J., Lockwood, K.: The value of reputation on eBay: A controlled experiment, Exp. Econ. 9, 79-101 (2006)
3. Fradkin, A., Grewal, E., Holtz, D., Pearson, M.: Bias and reciprocity in online reviews: Evidence from field experiments on Airbnb, Proc. 16th ACM Conference on Economics and Computation (2015).
4. Livan, G., Caccioli, F., Aste, T.: Excess reciprocity distorts reputation in online social networks, Sci. Rep. 7, 3551 (2017)

Graph analysis & word embedding help to perform user classification

Giuseppe Torrisi and Eric Fleury

Univ Lyon, ENS de Lyon, Inria, CNRS, UCBL, F69342, France,
giuseppetorrisi.gt@gmail.com, eric.fleury@inria.fr

1 Introduction

Twitter is an extremely rich social media engine used by a large spectrum of users: teenagers, businesses, politicians for campaigning or at the White House... Twitter is the place to express whatever you have on your mind to whoever wants to read it, with the only constraint being to say it in less than 140 characters. A key challenge is to capture the relation between people from the content they produce, in other words, not only the topics addressed, but also the way they are using the language gives a clue to predict likely social interactions. Our methodology is not based on #hashtags but encompasses all the content of tweets (words, emojis, abbreviations...). The underlying idea is that people that ”talk” similarly should have closely related interests. Our main contributions are the following: *(i)* we develop a generic framework for social media user classification which relies on the combined use of word embedding to capture language and topic similarity. *(ii)* we construct a proximity graph from these embeddings to perform user classification; *(iii)* we show that our framework can be instantiated and used with good results for Twitter; *(iv)* we validate our user classification by providing mention prediction analysis.

2 Methodology

The global framework is composed of four main phases: *(i)* The first one embeds the dictionary of words used in a large text corpora (all the tweets) into a low dimensional vector space Γ; *(ii)* from the tweets of a given user A, we compute a vector representation in Γ of A making use of the word-embedding vectors; *(iii)* we build a graph G of proximity between users from their coordinate in Γ; *(iv)* finally, we classify users by performing community detection on G and we forecast likely mentions.

From the tweet content, we tokenize and perform light sanitizing and use a class of neural network models that, from the unlabeled tweet corpus, produce a vector for each word/token in the corpus that encodes its semantic information. These vectors are useful since they effectively capture the semantic meanings of the words. We use Word2Vec [3], an unsupervised machine learning algorithm that embeds the dictionary of words into a low dimensional vector space Γ (typically of several hundred dimensions – 400 in our case). Word2Vec assigns to each token i a normalized vector $v(i) \in \Gamma$. In order to embed every user u into the space Γ, we compute the barycenter c_u of all words used by u and we re-normalized the result. In this way the resulting vector lays on the surface. We prove that such vector $c_u = \frac{\tilde{c}_u}{||\tilde{c}_u||}$ where $\tilde{c}_u = \frac{1}{k}\sum_{j=1}^{k=|\text{vocabulary of user } u|} v(j)$

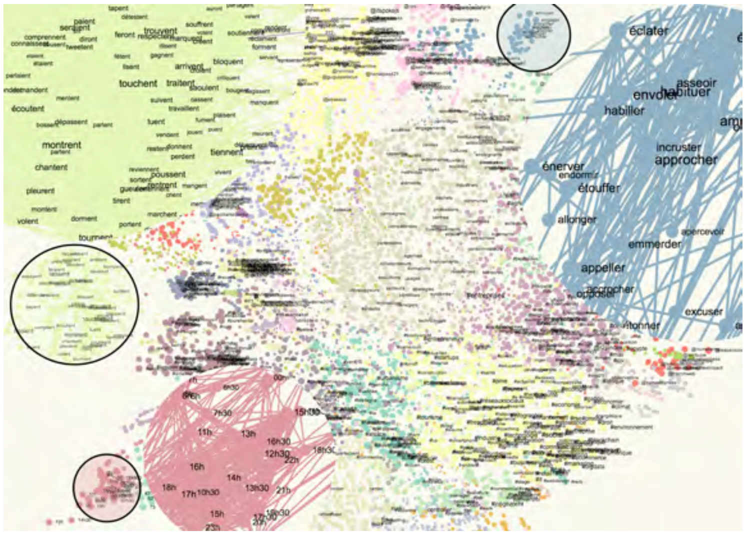


Fig. 1. Visualization of the proximity graph built from the word embedded in Γ and 3 zooms (blue, green, red) on specific communities. In blue the infinitive form of verbs from the 1st group (er); in green the conjugated form of the 3rd plural person; in red time related expressions (18h)

minimizes the distance to all other points (words used by the user u) on the surface of Γ. We then extract from the user embedded in the space Γ a proximity graph G (a user u is linked to its k closest neighbors v that are similar enough – the cosine between u and v is larger than $\alpha = 0.8$). We can then visualize the network G using ForceAtlas2 algorithm [2] and then apply a community detection algorithm [1].

3 Results

Corpus collection. Our dataset consists of a large data corpus collected from Twitter. Tweets may come with several types of metadata including information about the author's profile, the detected language, where and when the tweet was posted, etc. Specifically, we recorded 170 million tweets written in French, posted by 2.5 million users in the timezones GMT and GMT+1 over three years (between July 2014 to May 2017). These tweets were obtained via the Twitter powertrack API feeds provided by Datasift and Gnip with an access rate varying between $15-25\%$.

Word network. Note that if we skip the second step we can build a proximity graph G not for users but for the words (see Fig 1). Performing a community detection grasps basic grammatical elements, such as the infinitive form of verbs, plurals of nouns, numerical expressions... but also semantic topics (food, media...)

User network. Once user communities are computed, we label these communities according to the most representative words employed by the users of each community: we consider the words whose frequency in a given community is greater that the median plus two times the interquartile range (see Figure 2).

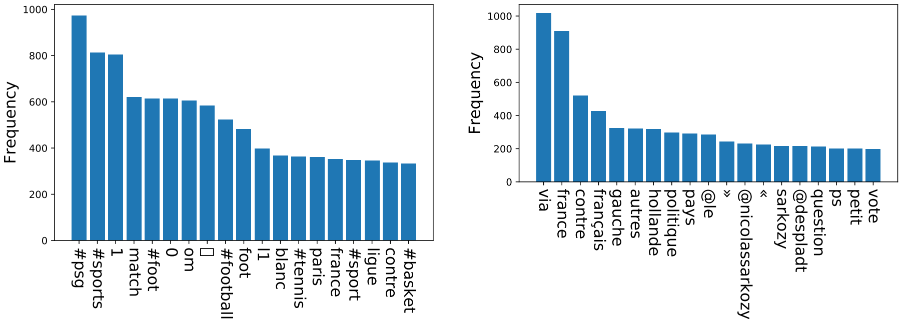


Fig. 2. Freq. distrib. of representative words appearing in a community of people talking about sports (left) and news (right). All words are related to the same topic, validating the fact that our graph construction is able to grasp real topic of discussions.

Mention prediction. To validate even further our user classification, we perform mention prediction. We expect that the distance between user-vectors infers social relations between people. We are able to predict mention between users and show the strong correlation between the success rate and their relative distance: the more the linguistic representation of users is similar, the more likely to mention each other.

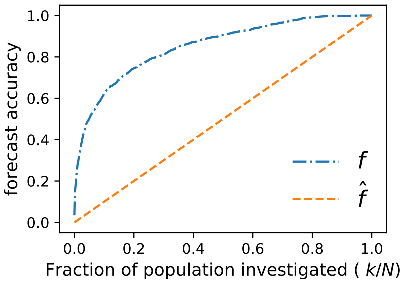


Fig. 3. Frequency f that, in a population of N users, one of k-closest neighbors is the most mentioned user (blue). The theoretical prediction $\hat{f}$ in the assumption of random connection (null model) is a linear trend (orange) . Our method outperforms the null model, especially at low k.

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10)
2. Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PloS one 9(6)
3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS 2013

Multiplayer ultimatum game on complex networks: the role of structural power

Fernando P. Santos[1,3], Francisco C. Santos[1,3], Ana Paiva[1], and Jorge M. Pacheco[2,3]

[1] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Tagusparque,
2744-016 Porto Salvo, Portugal
fernando.pedro@tecnico.ulisboa.pt,
[2] Centro de Biologia Molecular e Ambiental and Departamento de Matemática e Aplicações,
Universidade do Minho, 4710-057 Braga, Portugal
[3] ATP-Group, 1649-003 Lisbon Codex, Portugal

1 Introduction

The influence of fairness on human decision-making is often strong enough to overcome rationality and selfishness [1], posing a challenge to mathematical models that aim to justify the evolution of fair behaviour. While the dynamics of fairness in two-person interactions has been given significant attention, mostly in the context of Ultimatum Games (UG) [2], the challenge introduced by groups has not received the corresponding emphasis. Furthermore, the fact that individuals often participate in multiple groups makes it important to understand how the interdependence between different groups influences overall fairness. In this scenario, complex networks provide a key tool that enables specifying with whom each individual interacts and analyse how different interaction group assortments influence fairness.

Here we present a recent work [3] where we show that a single topological feature of social networks – which we call structural power (*SP*), measuring the average prevalence of one individual in the interaction groups of another – has a profound impact on the tendency of individuals to adopt fair strategies, in the context of Multiplayer Ultimatum Games (MUG) [4]. Increased fair outcomes are attained whenever *SP* is high, such that the networks that tie individuals allow them to meet the same partners in different groups, thus providing the opportunity to strongly influence each other.

1.1 Multiplayer Ultimatum Game and Structural Power

While in the 2-player UG a Proposer decides how to divide a given resource with a Responder and the game only yields payoff to the participants if the Responder accepts the proposal [2], in the N-player MUG proposals are made by one individual (the Proposer) to the remaining $N-1$ Responders, who must individually reject or accept the proposal [4]. Since individuals may act both as Proposers and Responders, we shall assume that each individual has a strategy characterised by two real numbers, p and q. The Proposer will try to split the endowment, offering p to the Responders. Each of the Responders will individually accept the offer made to the extent that his/her q-value is not larger than the p-value of the Proposer. Overall group acceptance will depend upon M, the minimum fraction of Responders that must accept the offer before it is valid. Consequently, if the fraction of individual acceptances stands below M, the offer will be rejected. Otherwise, the offer will be accepted. In this case, the Proposer will keep $1-p$ to himself and the group will share the remainder, that is, each Responder gets $p/(N-1)$. If the proposal is rejected, no one earns anything [4]. Values of $p=1-1/(N-1)$ and $q=1-1/(N-1)$, that provide Proposers and Responders with similar payoffs, are denoted fair.

We assume that MUG is played over an underlying network of contacts, considering the usual group formation process of multiplayer games on networks in which one node defines, together with his/her direct neighbours, a group [5]. This way, individuals may appear repeatedly in the interaction groups of others, which may provide increased structural power (SP) to some individuals over others. We define the *SP* of A over B as $SP_{A,B} = \frac{|I(A) \cap I(B)|}{|I(B)|}$, where $I(X)$ represents the groups in which individual X appears and $|I(X)|$ represents the number of groups in $I(X)$. One may note that, using the Kronecker $\delta_{A,B}$ to identify edges between A and B (e.g, 1 if an edge connects nodes A and B and 0 otherwise), denoting by $o_{A,B}$ (overlap) the number of common neighbours between A and B and by k_x the number of neighbours of X, then the *SP* of A over B is given by $SP_{A,B} = \frac{2\delta_{A,B}+o_{A,B}}{k_B+1}$. Intuitively, if one individual is a direct neighbour of other ($\delta_{A,B}=1$), they will meet in at least two groups, where each one will be the focal in each group. They will meet one more time for each of their $o_{A,B}$ common neighbours. If B has connectivity k_B, then this node participates in k_B+1 groups, providing the proper normalisation to $SP_{A,B}$ The average SP of one node is defined as $SP_A = |R(A)|^{-}1 \sum_{i \in R(A)} SP_{A,i}$, where $R(A)$ is the set of individuals reached by individual A, either directly or through a common neighbour.

To generate networks with pre-defined average *SP* and obtain the results portrayed in Fig. 1, we apply an optimisation algorithm to initially random networks. The random networks are generated by rewiring all the edges of a regular ring. Let us now assume that we want to build a network with average SP equal to sp_{max}. We re-organize the link structure of the initial network using a stochastic multi-step process such that, in each step, an edge of network is rewired at random (with no repeated edges allowed). The move is accepted if two criteria are met: 1) the resulting network remains connected and 2) the average SP of the resulting network (sp_t) increases (compared to the previous value) or passes the following stochastic criterion: a move in which *SP* decreases is accepted with probability $\lambda(sp_{max}-sp_t)$, where λ controls the probability of accepting an erroneous move. That means that the probability of accepting a rewire that decreases *SP* is lower as we get close to the desired *SP*. This is an optimisation feature similar in spirit to the well-known simulated annealing. We used $\lambda = 0.001$.

2 Results and Conclusion

We simulate the evolution of p and q in a population of size Z, much larger than the group size N. Initially, we equip individuals with values of p and q drawn from a discretised uniform probability distribution in the interval $[0,1]$ containing 101 values (discretised to the closer multiple of 0.01). The fitness F_i of an individual i of degree k is determined by the payoffs resulting from the game instances occurring in $k+1$ groups: one centred on her neighbourhood plus k others centred on each of her k neighbours.

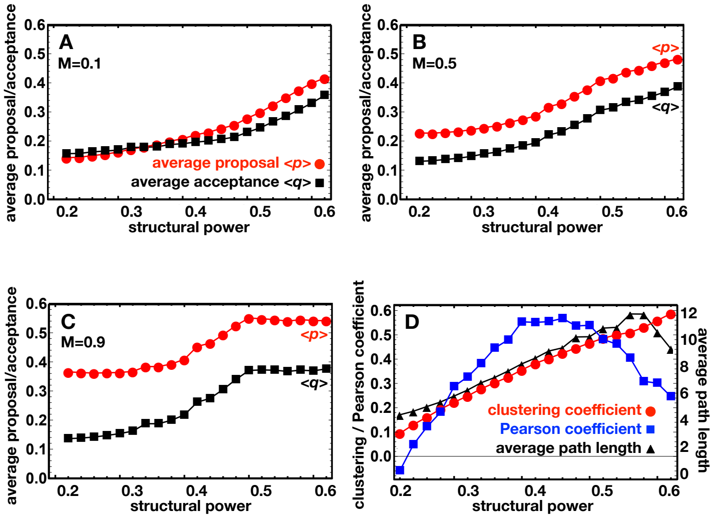


Fig. 1. For $M = 0.1$ (A), $M = 0.5$ (B) and $M = 0.9$ (C), networks with increased SP foster higher values of average p and q, i.e., strategies evolve to fairer levels. When generating networks with higher SP, other network metrics – such as clustering coefficient, average path length and degree-degree correlation – change (panel D), yet the increase in p and q is correlated with an increase in SP and not with the variation in other metrics [3].

Values of p and q evolve as individuals tend to imitate (i.e., copy p and q) those neighbours that obtain higher fitness values. Fig. 1 shows the average proposal p and acceptance threshold q that we obtain, as a function of the network SP. Clearly, high values of SP lead to higher average values of p and q, in which case individuals adopt fairer strategies. Our results suggest that a strong interdependence of groups taking part in collective decisions, here quantified by means of the SP, may be central in promoting seemingly paradoxical human features such as fairness.

Acknowledgments: This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references SFRH/BD/94736/2013, PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014, UID/BIA/04050/2013 and UID/CEC/50021/2013.

References

1. Fehr, E., Fischbacher, U.: The nature of human altruism. Nature 425(6960), 785 (2003)
2. Güth, W., Schmittberger, R., Schwarze, B.: An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization 3(4), 367–388 (1982)
3. Santos, F.P., Pacheco, J.M., Paiva, A., Santos, F.C.: Structural power and the evolution of collective fairness in social networks. PloS ONE 12(4), e0175687 (2017)
4. Santos, F.P., Santos, F.C., Paiva, A., Pacheco, J.M.: Evolutionary dynamics of group fairness. Journal of Theoretical Biology 378, 96–102 (2015)
5. Santos, F.C., Santos, M.D., Pacheco, J.M.: Social diversity promotes the emergence of cooperation in public goods games. Nature 454(7201), 213 (2008)

MATI: An Efficient Algorithm for Influence Maximization in Social Networks

Maria-Evgenia G. Rossi[1], Bowen Shi[1], Nikolaos Tziortziotis[1],
Fragkiskos D. Malliaros[2], Christos Giatsidis[1], and Michalis Vazirgiannis[1]

[1] École Polytechnique, France
[2] Center for Visual Computing, CentraleSupélec and Inria, France and UC San Diego, USA
E-mail: maria.rossi@polytechnique.edu

1 Introduction

Influence maximization (IM) has attracted a lot of attention due to its numerous applications, including diffusion of social movements, the spread of news, viral marketing and outbreak of diseases. The problem can formally be described as follows: given a *social network* where the relations among users are revealed, a *diffusion model* that simulates how information propagates through the network and a parameter k, the goal is to locate those k users that maximize the spread of influence. Kempe et al. [3] formulated the problem in the aforementioned manner while adopting two diffusion models borrowed from mathematical sociology: the *Linear Threshold* (LT) and the *Independent Cascade* (IC) model. According to both, at any discrete time step a user can be either active or inactive and the information propagates until no more users can be activated.

In this study, we propose MATI, an efficient IM algorithm under both the LT and IC models. By taking advantage of the possible paths that are created in each node's neighborhood, we have designed an algorithm that succeeds in locating the users that can maximize the influence in a social network while also being scalable for large datasets. In order to limit the computation of the possible paths and the respective probabilities of them being "active", we use a pruning threshold θ that reduces the running time but also the accuracy of the influence computation. Extensive experiments show that MATI has competitive performance when compared with the baseline methods both in terms of influence and computation time. Due to space limitations we present only the respective methodology and results for the MATI algorithm under the LT model.

2 MATrix Influence (MATI) algorithm

A social network is typically modeled as a directed graph $G = (V, E)$, consisting of $|V|$ users represented as nodes and $|E|$ edges reflecting the relationship between users. We assume that $\mathscr{T}(u) = \{\tau_1, \tau_2, \dots, \tau_M\}$ represents the set of all possible paths that exist in the graph starting from node u and leading to "leaf" nodes. Each path τ_i consists of a sequence of nodes: $\tau_i = \{n_{i1}, n_{i2}, \dots, n_{iN}\}$. Let $p^{\tau}_{\ell,\ell+1}$, $1 \leq \ell \leq N-1$, represent the influence weight (probability) between two successive nodes in path τ. Then $\mathscr{F}(\tau_i) = \{f_{i1}, f_{i2}, \dots, f_{iN}\}$ represents the probability path for every path τ_i starting from node u to be *a*ctive. Each f_{ij} is equal to $\prod_{\ell=1}^{j-1} p^{\tau_i}_{\ell,\ell+1}$ if $j \geq 1$, and 1 otherwise.

Algorithm 1 MatiLT

```
Input: $G = (V, E)$, $k$                          ▷ $k$: budget (number of seed nodes)
Initialize: $S = \emptyset$
$\mathscr{A}, \Omega = $ CalcStatsLT$(G)$
$Q = $ CalcInf$(\mathscr{A}, V)$                   ▷ $Q$: CELF queue [4]
for $i = 1$ to $k$ do
    $s, \sigma(s) = Q.top()$
    $S = S \cup s$
    $U = V \backslash S$
    for each $u \in U$ do
        $\sigma(u) = Q(u)$
        for each $v \in S$ do
            $\sigma(u) -= \Omega(v, u)$
            $\sigma(u) -= \Omega(u, v)$
        end for
        $Q.add((u, \sigma(u)))$
    end for
end for
return $S$
```

The *forward cumulative influence* $\Omega(u,v)$ corresponds to the influence of node u to v and to the nodes that can be found right after v in the paths $\mathscr{T}(u)$ of node u.

Goyal et al. [2] showed that the spread of a set S of nodes is the sum of the spread of each individual node $u \in S$ on the subgraphs induced by the set $V - S + u$:

$$\sigma(S) = \sum_{u \in S} \sigma^{V-S+u}(u), \tag{1}$$

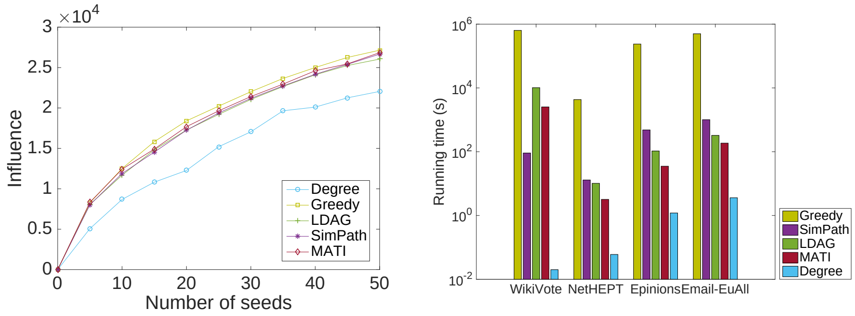
where $\sigma^{V-S+u}(u)$ denotes the total influence of u in the subgraph induced by $V - S + u$. Similar to [2], we write $V - S$ to denote the difference of sets V and S, $V \setminus S$, and $V - S + u$ to denote $((V \setminus S) \cup \{u\})$.

Theorem 1 constitutes the core of the MATI algorithm under the LT model. Actually, it is used for the calculation of the influence gain after the addition of a node x to a set of nodes S.

Theorem 1 *Under the LT model, to calculate the influence after adding a node x to a set of nodes S, one has to subtract from the sum of the individual spread of S and x the forward cumulative influence Ω of all the nodes that belong to set S which contain node x in paths connecting the latter to nodes in set S. That is,*

$$\sigma(S+x) = \sigma(S) + \sigma(x) - \sum_{y \in S} \Omega(x,y) - \sum_{y \in S} \Omega(y,x).$$

Alg. 1 shows the complete structure of MATI algorithm under the LT model. CalcStatsLT computes $\mathscr{A}$ (i.e., $\mathscr{A}(S,u)$ is the probability the single node u to be activated (influenced) by S) and Ω, and CalcInf returns the influence of all nodes $v \in V$.



(a) (b)

Fig. 1: (a) Influence spread in number of nodes under the LT model for the EPINIONS dataset. (b) Comparison of running times in seconds.

3 Empirical Analysis

We have conducted experiments in real-world datasets in order to evaluate the performance of the MATI algorithm and compare it to state-of-the-art influence maximization algorithms on the quality of results and efficiency. We have used four publicly available graph datasets[3]: NETHEPT, WIKIVOTE, EPINIONS and EMAIL-EUALL and compared the respective results with those of four baseline algorithms: i) *Degree* which considers high-degree nodes as influential [3], ii) *Greedy* for which following the literature [3], we run $10,000$ Monte Carlo (MC) simulations to estimate the spread of any seed set, iii) *LDAG* algorithm using locality properties as proposed in [1] and iv) *SimPath* algorithm proposed in [2]. The threshold θ for the MATI algorithm is set to 0.0001.

The quality of the seed sets obtained by different algorithms is evaluated based on the expected spread measured in number of nodes (Fig. 1a). Due to space limitations we only present the respective results for the EPINIONS dataset. The seed sets obtained via MATI are quite competitive in quality compared to those of the Greedy, LDAG and SimPath algorithms. For all four datasets, the influence loss for up to 50 seeds is less than 2%. Figure 1b reports the execution time required by various algorithms for the LT model. In all cases, MATI is faster than the Greedy and LDAG algorithms. In all datasets except WikiVote, MATI also performs slightly better that SimPath.

References

1. Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model. In: ICDM (2010)
2. Goyal, A., Lu, W., Lakshmanan, L.V.S.: Simpath: An efficient algorithm for influence maximization under the linear threshold model. In: ICDM (2011)
3. Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD (2003)
4. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: KDD (2007)

[3] https://snap.stanford.edu/data/index.html

Team success in the iGEM scientific competition

Marc Santolini[1,2], Abhijeet Krishna, Christos Ellinas, Leo Blondel, Thomas E. Landrain, and Albert-Laszló Barabásí[1,2]

[1] Northeastern University, Boston MA
[2] Harvard Medical School, Boston MA
m.santolini@neu.edu

Abstract. This work investigates criteria of performance and success of teams in a scientific context. We leverage laboratory notebooks edited on wiki websites by student teams participating to the international Genetically Engineered Machines (iGEM) synthetic biology competition to uncover what features of team work best predict short term quality (medals, prizes) and long term impact (how the biological parts that teams engineer are re-used by other teams). This represents a large-scale dataset of 2,000 teams over 10 years, with an average 10 students per team, providing an unprecedented insight into the making of science.

1 Introduction

Recently, the literature has bloomed in large scale analyses of science as an object of study, paving the way for the ”Science of science”. Pervasive to this literature is the use of networks. Indeed, research is a collective phenomenon where any finding relies on previous knowledge elaborated by others [1]. The co-authorship produces a collaboration network informative on the social biases of research [2, 3], and the citation network allows to measure the impact and spread of new concepts [4].

While there is a trove of large scale datasets relative to the outputs of science, much less can be said with regards to the making of science *in situ*, in the laboratory. At the qualitative level, social scientists have been investigating this question decades ago, with early work by Latour and Woolgar [5] exhibiting the anthropological aspects of making science. Yet, such investigations have been lacking a quantitative counterpart, in part due to technical limitations. The necessary toolkit is nonetheless ready. For example, team metrics and their relation to team success have been measured in collaborative coding [6], in the artistic setup of Broadway musical [3], or in private organisations [7, 8]. Here we explore their role in the context of scientific production.

We leverage the iGEM Competition[3] of Synthetic Biology. For over 10 years, iGEM has been encouraging students to work together to solve real-world challenges by building genetically engineered biological systems with standard, interchangeable parts or BioBricks. Student teams design, build and test their projects over the summer and gather to present their work and compete at the annual Jamboree. A condition of participation to iGEM is that teams document their progress

[3] http://igem.org/Main_Page

and results on an open wiki website[4]. Given the underlying structure of wikis, it is possible to know which team member has edited which part of the wiki at what time. Finally, teams are awarded medals and special prizes (short term impact), and the BioBricks that they engineer can be later re-used by other teams in later years (long term impact). In this work, we investigate how features of team organization (obtained through the wiki) affect team success (medals, prizes etc) in this scientific context.

2 Results

We extracted team information at multiple levels, as shown in Figure 1. First, we built a scraper to extract the wiki history and content for 1,551 teams, information that was used to build internal team interaction networks. For each team, a bipartite network was constructed between the wiki editors (the team students) and the sections edited. Team networks were then reconstructed by projecting the bipartite network on the user space, counting the number of wiki subsections co-edited by any two students of a team. The obtained number was compared to the expected co-edition resulting from a hypergeometric distribution and a Z-score was computed. Finally, edges with $Z > 2$ were deemed significant and kept for further analyses. Teams also collaborate with one another, and we extracted for each year the team collaboration network. Teams produce BioBricks, and we extracted the number of BioBricks produced and their re-use. Finally, success measures were collected, consisting of the type of medal (None, Bronze, Silver or Gold), number of special prizes, being a Finalist and being a Winner of the competition.

Analysis of the data showed higher productivity per capita (number of edits, number of sections edited, number of bioparts produced) and larger time invested in the project for teams winning higher quality medals. Moreover, we observed that teams with higher network density, largest connected component, and a leader with high degree were significantly more successful in the competition. Finally, we built a classifier combining these different features together by fitting a logistic regression model with cross-validation. The classifier was able to predict winning teams with an Area under the ROC Curve $AUC = 0.8$, showing the high degree of relevance of the captured features with the observed outcome.

3 Discussion and future prospects

The iGEM competition offers a model system to understand the making of science in a controlled context. Here we have leveraged the publicly available data from the competition by making use of the open wiki websites. Future prospects consist in exploring the finer-grained judging data currently being obtained from the iGEM HQ. For each team, 6 judges rate 60 criteria from 1 to 5, quantifying project creativity and quality[5]. By accessing this data, we will have an unprecedented insight into how science quality is assessed, the variability of the assessment between judges, as well as what features correlate with more subjective metrics such as project creativity.

[4] To see an example wiki: `http://2016.igem.org/Team:LMU-TUM_Munich`

[5] `http://2017.igem.org/Judging/Rubric`

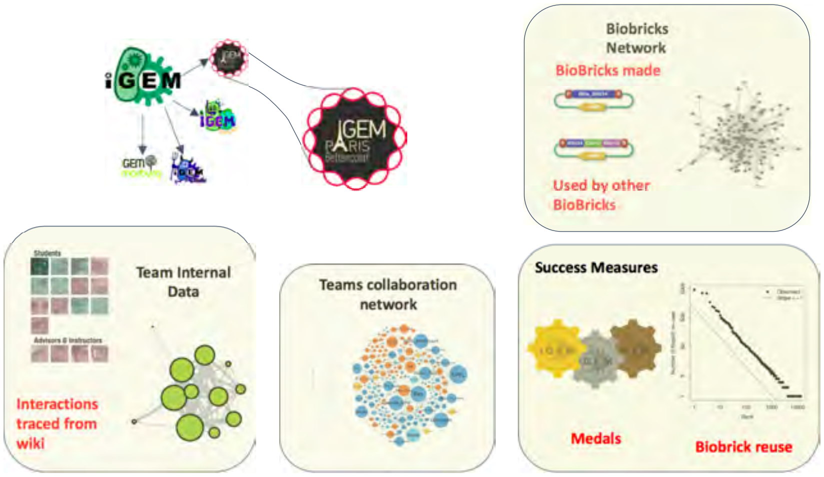


Fig. 1. Overview of the dataset. Over 10 years, 2,000+ teams have participated to the iGEM competition. Team internal networks are reconstructed from wiki notebooks co-edition. Teams collaborate with one another, forming a collaboration network. They produce BioBricks by combining previously made BioBricks or engineering new ones. Finally, team success is determined by the prizes and medals they receive, as well as their BioBricks re-use by other teams.

References

1. Sinatra,R., Deville,P., Szell,M., Wang,D. and Barabasi,A.-L. (2015) A century of physics. Nature Physics, 11, 791–796.
2. Shen,H.-W. and Barabasi,A.-L. (2014) Collective credit allocation in science. Proc Natl Acad Sci USA, 111, 12325–12330.
3. Guimer,R., Uzzi,B., Spiro,J. and Amaral,L.A.N. (2005) Team assembly mechanisms determine collaboration network structure and team performance. Science, 308, 697–702.
4. Uzzi,B., Mukherjee,S., Stringer,M. and Jones,B. (2013) Atypical Combinations and Scientific Impact. Science, 342, 468–472.
5. Latour, B., Woolgar, S (1979). Laboratory Life: The Social Construction of Scientific Facts, Beverly Hills, Sage Publications, 1979. (ISBN 0803909934) ; rd. Princeton, Princeton University Press, 1986
6. Klug,M. and Bagrow,J.P. (2016) Understanding the group dynamics and success of teams. R. Soc. open sci., 3, 160007.
7. Pentland,A. (2012) The new science of building great teams. Harvard Business Review.
8. Watts, D. (2016), The Organizational Spectroscope (`https://medium.com/@duncanjwatts/the-organizational-spectroscope-7f9f239a897c`)

Part II

Biological Networks

Modular decomposition of protein structure using community detection

William P. Grant[1] and Sebastian E. Ahnert[1,2]

[1] Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, UK.
wpg23@cam.ac.uk
[2] Sainsbury Laboratory, University of Cambridge, UK.

1 Introduction

Proteins are key to all cellular life, as the components which carry out almost all functions within the cell. Analysing the set of solved protein structures (as stored in the Protein Data Bank [6]) may reveal common organisational and evolutionary principles, and assist in identifying conserved regions of protein structure.

In this work, a given protein structure is abstracted into an undirected, weighted network, in which either the atoms or the amino acids of the protein correspond to vertices. Edges are generated using a simple distance threshold. This allows the pattern of bonding within the protein to be explored using existing network analysis methods. This method has previously shown promise in identifying key amino acids for allostery [1] and for protein stability [3].

Using community detection (specifically the Infomap algorithm [7]), highly intra-connected regions of the network can be found. These are then mapped back onto regions of the protein's structure. This analysis can be carried out across large numbers of structures, in order to find regions of structure that may be repeated across distinct proteins. The number and arrangement of the resulting communities can also be investigated, as a potential description of the protein's topology.

2 Results

The community structure at a given length scale can be compared with existing annotations on the protein structure by treating the community structure and the existing annotation as two descriptors of the protein sequence. The Jaccard index can then be used to compare the "generated" and "expected" arrays [5]. In this work the PFAM sequence domains [4] are compared to the community structure. A modified version of the Jaccard index is used, which does not penalise community structure generated outside the regions of PFAM classification. The significance of the resulting match can be found using the z-score, with null models preserving the total number of communities, and the number of boundaries between communities. Figure 1 shows significant agreement between the regions of the protein corresponding to PFAM domains, and the regions corresponding to communities.

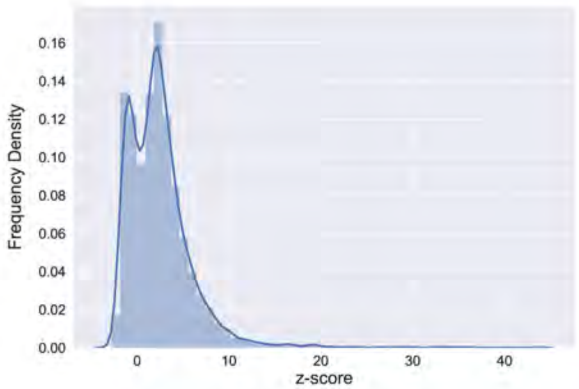


Fig. 1. Histogram showing the z-score for the Jaccard index between the generated community structure, and the PFAM domains, for $\sim$ 1000 test proteins, showing that in many cases the agreement is extremely significant. Here protein residue networks are used, with a distance threshold of four times the respective covalent radii.

In addition to the communities' intrinsic value as globular (possibly conserved) subunits, the arrangement of the communities may give insight into the protein's topology. This was investigated by generating condensation networks in which the nodes of the network correspond to communities on the original network, linked if an edge links the original communities. Grouping proteins according to their condensation network may reveal common domain architecture. Figure 2 shows the set of condensation networks for the 12,000 proteins studied so far. The proteins corresponding to each condensation network are displayed interactively at tcm.phy.cam.ac.uk/~wpg23. Enriched Gene Ontology (GO) terms [2], detailing potential functional classes, are also displayed.

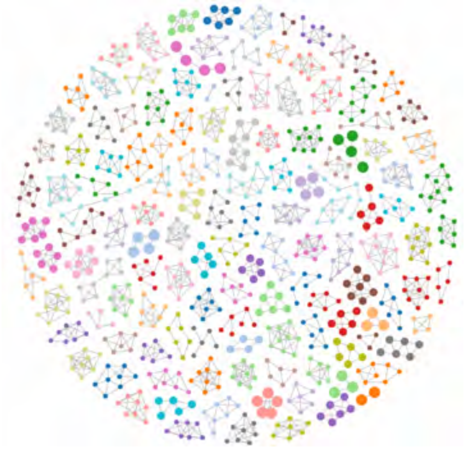


Fig. 2. The condensation networks found from the domain-level Infomap community structure of residue contact networks, from a dataset of 6,000 proteins. 129 condensation networks are shown, with networks corresponding to only one protein excluded. Node size indicates the number of proteins corresponding to each condensation network. The member proteins and associated GO terms for each network can be found at tcm.phy.cam.ac.uk/~wpg23.

References

1. Amor, B.R., Schaub, M.T., Yaliraki, S.N., Barahona, M.: Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. Nature Communications 7 (2016)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature Genetics 25(1), 25–29 (2000)
3. Delmotte, A., Tate, E.W., Yaliraki, S.N., Barahona, M.: Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin–myosin light chain interaction. Physical Biology 8(5), 055010 (2011)
4. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A.: The PFAM protein families database: Towards a more sustainable future. Nucleic Acids Research 44(D1), D279–D285 (Jan 4 2016)
5. Fortunato, S., Hric, D.: Community detection in networks: A user guide. Physics Reports 659, 1–44 (2016)
6. Rose, P.W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z., et al.: The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. Nucleic Acids Research 45, D271–D281 (2017)
7. Rosvall, M., Bergstrom, C.T.: Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PLOS ONE 6(4), 1–10 (04 2011)

Reconfiguration of Protein Interaction Networks during Nematode Development

Florian Klimm[1], Charlotte M. Deane[1], Jonny Wray[2], and Mason A. Porter[3]

[1] University of Oxford Woodstock Rd, Oxford OX2 6GG
[2] e-Therapeutics Plc 17 Blenheim Office Park, Long Hanborough, OX2 98LN
[3] University of California at Los Angeles Box 951555 Los Angeles, CA 90095-1555

1 Introduction

Protein interaction networks (PINs) allow the representation and analysis of biological processes in cells. Because cells are dynamic and adaptive, these processes change over time. One example of adaptive regulation is the change of gene expression, which may occur at very different time scales [1]: responses to environmental signals take minutes [2], and developmental changes take days in *C. elegans* [3] and years in humans [4]. This change in protein expression results in an altered protein abundance in an organism.

Thus far, research has focused either on the static PIN analysis or the temporal nature of gene expression. By analysing temporal PINs using multilayer networks [5], we want to link these efforts. The construction and analysis of temporal PINs gives insights into how proteins, individually and in their entirety, change their biological functions. In our investigation, we find that modular structure in the roundworm *C. elegans*' PIN changes during development. Using gene ontology (GO) terms, we connect this structural change with a reorganisation of biological functions. To our knowledge, our results represent the first direct identification of dynamic modular structure in PINs, despite having been hypothesised more than a decade ago [6].

2 Data sets

We use the mutlilayer network constructed in [7]. It consists of a total of $N = 4,792$ proteins. Interactions between them are aggregated from BioGrid [8] and other protein interaction databases. This gives a monolayer network of all protein interactions, as shown in Fig. 1. Each layer is then constructed as a subnetwork consisting of all proteins expressed at that developmental stage and all interactions between them. The gene expression information for six developmental stages (blastula, gastrula, embryo, nematode, prime adult, and life cycle) is extracted from the Bgee repository [9]. The layers consist of a variable number of nodes, ranging from 2,848 in the gastrula stage to 4,755 in the nematode stage.

After the construction of the layers, we connect them with interlayer edges of different strength ω. In this abstract, we illustrate results only for $\omega = 0.1$, but we will present results for multiple values in the oral presentation. We exclusively connect two nodes in layers of successive developmental stages and if they represent the same protein.

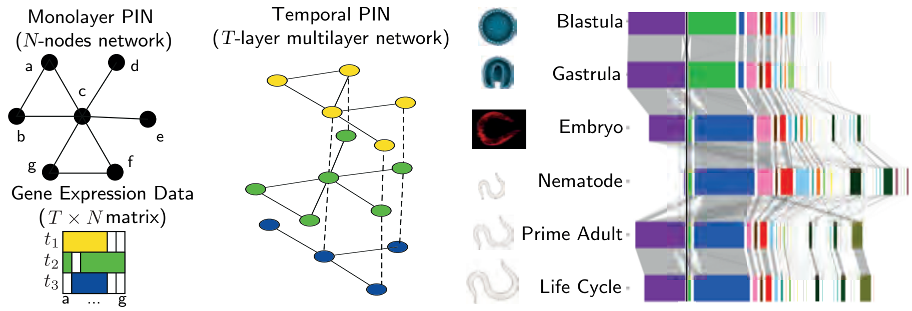


Fig. 1. Left: Temporal PIN links temporal gene expression with static PIN. Example given for $T = 3$ time points and $N = 7$ nodes. Right: Alluvial plots of the developmental modular structure of *C. elegans*' PIN . Rectangles represent modules of nodes with their width indicating the module's size. The most left rectangle in purple represent all proteins that are not expressed in a given developmental stage. The width of the gray lines indicate the overlap between modules in temporally adjacent layers and thus give a strength of transition from one developmental stage to the next.

3 Results

Because the functionality of biological processes change during the development, we suspect that modular structure also changes during development. To test this hypothesis, we detect and analyse the modular structure in the developmental PIN of *C. elegans*. We use GenLouvain [10], a modularity optimisation method suited for multilayer networks, for the community detection.

The detected modular structure (see Fig. 1) consists of two facets: The network is organised in modules inside each layer and modules change over time (i.e., across layers). The modular structure inside each layer gives an indication of the functional organisation of the proteins at a given developmental stage. The modules vary in size and the number of modules in each layer ranges from eleven to more than twenty.

We use GO enrichment to test whether the detected structural modules consist of proteins with a mutual function. We use a significance level of 0.05 and use Bonferroni correction to take into account the problem of multiple comparisons, because we test the enrichment of more than 2×10^6 GO terms. The large modules tend to show enrichment for fairly broad terms, such as 'protein binding'; and the smaller modules, show more specific terms, such as 'embryo development' and 'proteasome complex'. This is consistent with earlier results, which show that, different GO terms tend to be enriched at different module sizes [11].

The modules are often enrichmed for many different GO terms at the same time. For example, module 9 (marked in light green) of blastula stage is enriched for approximately 50 terms. Amongst them are many terms that reflect different developmental processes like 'embryo development', 'larval development', 'hermaphrodite genitalia development', and 'reproduction of symbiont in host'.

We find that some parts of modular structure stay similar during the developmental, whereas others undergo considerable change. The reconfiguration of modular structure over time can give additional insights into the adaptive function of a cell. To give one example, we focus on module 5 (marked in red) of the nematode stage. GO enrichment indicates that its dominant function is ‘embryo development’. Its members are in three different modules at the next stage ‘prime adult’. This suggests that this function may adapt and is now distributed across those three modules. To investigate this further, we separately examine the GO enrichment for each of these three groups of nodes that change their module. We find that all three of them have enriched ‘embryo development’, but one of them has a much stronger enrichment for ‘ubiquitin-independent protein catabolic process via the multivesicular body sorting pathway’. This suggests that multivesicular body sorting pathway is connected to the embryo development.

Summary. We represent the PIN of the nematode *C. elegans* during development as a mutlilayer network. By investigating modular structure in this network, we detect a partial reconfiguration of its communities with development. Further comparison of the modular structure with gene ontology annotations hints at biological functions of the modules of proteins. By examing the structural change of modules from one developmental stage to the next we are able to detect modules that break apart or combine. This hints at functional change such as a strengthening of subfunctions.

References

1. Luis López-Maury, Samuel Marguerat, and Jürg Bähler. Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, 9(8):583, 2008.
2. Nir Yosef and Aviv Regev. Impulse control: Temporal dynamics in gene transcription. *Cell*, 144(6):886–896, 2011.
3. George L Sutphin and Matt Kaeberlein. Measuring *Caenorhabditis elegans* life span on solid media. *Journal of Visualized Experiments*, (27), 2009.
4. Urie Bronfenbrenner. The ecology of human development: Experiments by nature and design. *American Psychologist*, 32:513–531, 1979.
5. Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.
6. Jing-Dong J Han, Nicolas Bertin, Hao Tong, Debra S Goldberg, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88, 2004.
7. Pramod Shinde and Sarika Jalan. A multilayer protein–protein interaction network analysis of different life stages in *Caenorhabditis elegans*. *EPL (Europhysics Letters)*, 112(5):58001, 2015.
8. Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: A general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl_1):D535–D539, 2006.
9. Frederic Bastian, Gilles Parmentier, Julien Roux, Sebastien Moretti, Vincent Laudet, and Marc Robinson-Rechavi. BGee: Integrating and comparing heterogeneous transcriptome data among species. In *Data Integration in the Life Sciences*, pages 124–131. Springer, 2008.
10. Lucas G. S. Jeub, Marya Bazzi, Inderjit S. Jutla, and Peter J. Mucha. A generalized louvain method for community detection implemented in MATLAB, version 2.1. http://netwiki.amath.unc.edu/GenLouvain, 2011-2014.
11. Anna CF Lewis, Nick S Jones, Mason A Porter, and Charlotte M Deane. The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology*, 4(1):100, 2010.

Using assortativity and other network properties to unravel the organization of chromatin in the nucleus

Vera Pancaldi[1], Enrique Carrillo-de-Santa-Pau[2], David Juan[3], Alfonso Valencia[1], and Daniel Rico[4]

[1] Barcelona Supercomputing Centre (BSC-CNS) Barcelona 08034, Spain
vera.pancaldi@bsc.es
[2] Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain
[3] Institut de Biologia Evolutiva, Consejo Superior de Investigaciones Cientificas, Universitat Pompeu Fabra, Parc de Recerca Biomedica de Barcelona, Barcelona, 08003, Spain
[4] Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK

1 Introduction

In recent years a number of biological networks have been mapped and have attracted the attention of network scientists. Most recently, technical advances in mapping chromatin 3D interactions inside the nucleus have allowed us to start studying networks of genomic elements. The conformation of the genome as well as the presence of specific modifications on the DNA (epigenomic marks) and binding of proteins (transcription factors) to the DNA are fundamental in regulating the expression of genes, that ultimately shapes cells and their behaviour. Many datasets are available detailing the location of these epigenetic features along the genome, but only recently we have been able to put these data into the context of the 3D chromatin structure. Interactions have been shown to exist between different genes (linking gene promoters between them) and also between genes and their regulatory regions, spanning varying genomic distances and even uniting different chromosomes. We apply network methods to integrate these different data types and to unveil chromatin organizing principles.

2 Results

Here we present a method which uses assortativity of specific epigenomic marks (including chromatin modification and transcription factor binding sites) to identify which factors are mostly responsible for the 3D organization of chromatin. Assortativity has been used in the past to identify networks in which preferential contacts are established between nodes that share certain properties (for example ethnicity was shown to be assortative in school networks and happiness on twitter networks [3, 1]. In our case, we suggest that proteins whose binding peaks are assortative in the chromatin network are likely major players in establishing the 3D contacts.

Assortativity identifies main structural determinants of chromatin 3D structure We used close to 80 epigenomic features mapped along the genome in mouse embryonic stem cells (as a binary variable defined in 200bp windows), a very well characterised biological system [2], and combined these data with networks of chromatin interaction in the same cell type determined using Promoter Capture HiC (denoted as PCHiC networks here) [**?**]. In these networks each node represents a genomic fragment (mean length 5kb) and an edge is present if two chromatin fragments are found to be interacting in 3D [5]. For each fragment, we calculated the proportion of fragment covered by each feature and called the average of this value across the genome abundance. We then calculated the assortativity of the abundance of each genomic feature (ChAs) and plotted it against the abundance. The features with high assortativity despite having moderate abundance are likely to be important determinants of 3D genome organization (Fig.ChAs).

We find many features with significantly higher ChAs than expected at random and identify binding sites of proteins that are known to be involved in chromatin organization (Polycomb) as the most assortative features [4] (Fig.ChAs).

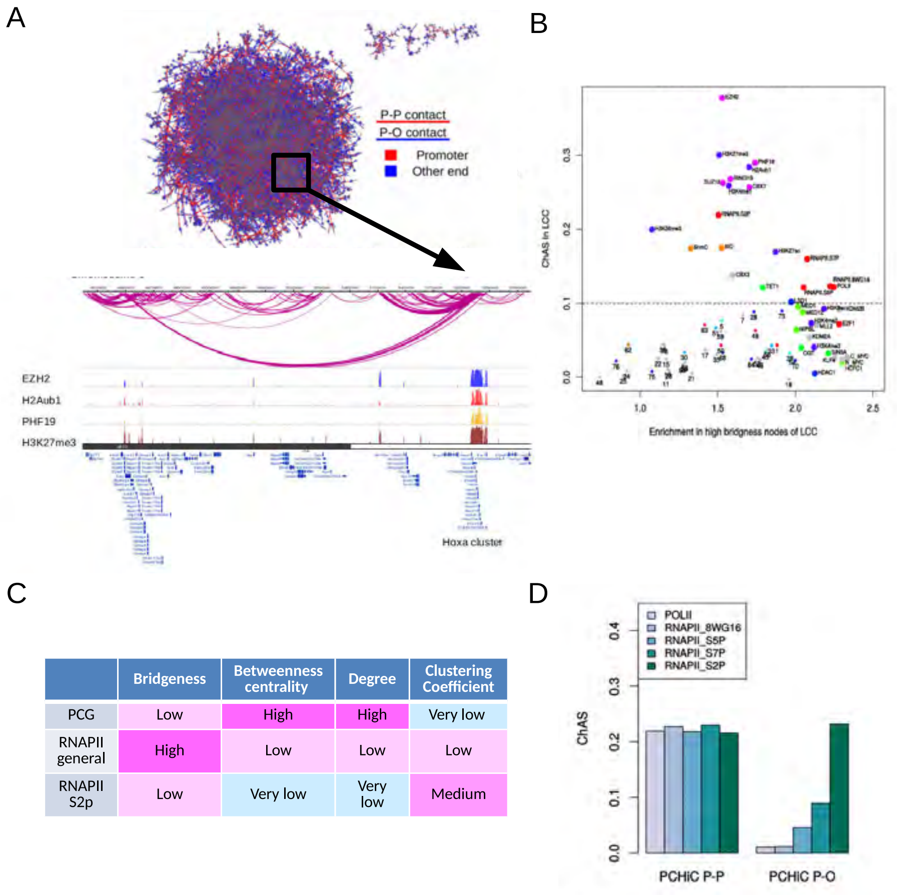
Characterising interactions between promoters and regulatory regions The PCHiC network can be split into 2 subnetworks: 1) the network of contacts between genes (promoter-promoter contacts, PP) and 2) the network uniting genes with their regulatory regions (promoter- other end contacts, PO). Due to the chosen experimental technique, we do not have any coverage of the contacts not involving promoters. Interestingly we observe pronounced differences in the ChAs values of some features in these two subnetworks, suggesting that different factors mediate different types of contacts.

Importantly, we notice differences in the ChAs value of different forms of RNA polymerase in the PP and PO networks. RNA polymerase is a protein involved in transcription and it is found in various states which are associated with different phases of RNA transcript production, from a paused state to the actively elongation state. Active elongation is thought to depend on contacts betwee regulatory regions and the gene itself. Whereas in the PP subnetwork all forms of RNA polymerase have equal ChAs, the ChAs of non-elongating forms is reduced in PO contacts. This suggests that promoters occupied by polymerase are in preferential contact between them but preferential contacts with regulatory regions are only established for the actively elongating form.

Network properties associated to different chromatin features Performing a topological analysis and defining overlapping modules on the network using Moduland [6], we further showed that characteristic topological properties of nodes occupied by different epigenomic features. For example, nodes occupied by the main organizers of 3D structure (Polycomb proteins) are found to have high betweenness centrality but low bridgeness (number of overlapping modules that a node belongs to), suggesting that Polycomb is not involved in contacts between different chromatin modules. RNA polymerase is found to have high betweenness and high bridgeness, suggesting that it mediates contacts between different chromatin modules. On the contrary, the elongating form of RNA polymerase is enriched in nodes with low betweenness and centrality, suggesting that they would remain more peripheral.

3 Conclusion

These findings recapitulate known biology, as far as the major role of polycomb is concerned, and offer new insights about the important role of elongation in contacts between regulatory elements and their target genes. The framework established is very general and enables studying the 3D organization and role of different genomic features.



	Bridgeness	Betweenness centrality	Degree	Clustering Coefficient
PCG	Low	High	High	Very low
RNAPII general	High	Low	Low	Low
RNAPII S2p	Low	Very low	Very low	Medium

Fig. 1. Graphical summary of the network analysis. A) Chromatin contact network showing larger connected components and mapping of genomic features on the network nodes. B) Chromatin Assortativity (ChAs) of features versus bridgeness of nodes occupied by those features. C) Summary table of network statistics for nodes with particular chromatin features. D) ChAs values for different states of RNA pol II in PP and PO subnetworks.

References

1. Bollen, J., Gonalves, B., Ruan, G., Mao, H.: Happiness is assortative in online social networks. Artificial Life 17(3), 237–251 (2011), pMID: 21554117

2. Juan, D., Perner, J., Carrillo de Santa Pau, E., Marsili, S., Ochoa, D., Chung, H.R., Vingron, M., Rico, D., Valencia, A.: Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs. Cell reports 14(5), 1246–57 (feb 2016)
3. Newman, M.E.J.: Assortative Mixing in Networks. Physical Review Letters 89(20), 208701 (oct 2002)
4. Pancaldi, V., Carrillo-de Santa-Pau, E., Javierre, B.M., Juan, D., Fraser, P., Spivakov, M., Valencia, A., Rico, D.: Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biology 17(1), 152 (dec 2016)
5. Sandhu, K.S., Li, G., Poh, H.M., Quek, Y.L.K., Sia, Y.Y., Peh, S.Q., Mulawadi, F.H., Lim, J., Sikic, M., Menghi, F., Thalamuthu, A., Sung, W.K., Ruan, X., Fullwood, M.J., Liu, E., Csermely, P., Ruan, Y.: Large-scale functional organization of long-range chromatin interaction networks. Cell reports 2(5), 1207–19 (nov 2012)
6. Szalay-Beko, M., Palotai, R., Szappanos, B., Kovács, I.A., Papp, B., Csermely, P.: ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping network modules and community centrality. Bioinformatics (Oxford, England) 28(16), 2202–4 (aug 2012)

Hypergraphlets Give Insight into Multi-Scale Organisation of Molecular Networks

Thomas Gaudelet[1], Noel Malod Dognin[1], Jose Lugo-Martinez[2], Predrag Radivojac[2], and Natasa Pzrulj[1]

[1] University College London, Department of Computer Science, London, WC1E 6BT, United-Kingdom,
n.pzrulj@ucl.ac.uk
[2] Indiana University, Department of Computer Science, Bloomington, Indiana 47405, U.S.A.

1 Introduction

Deciphering the complex patterns of interactions between macro-molecules in a cell is of crucial importance in systems biology. Graph theory offers abstractions necessary to represent molecular interactions and to study them from a mathematical perspective, aiming to uncover patterns in their interconnectedness that could be linked to biological reality. Notably, binary graphs have been widely used to capture the interactions between pairs of genes. Graphlets [2, 3] and their statistics have been introduced and used as an underlying to study these binary representations with objectives such as: comparing biological networks, uncovering their functional organisation principles, relating the wiring patterns of genes in molecular networks with their biological functions [4, 8].

However, molecules often do not interact solely in a pairwise fashion and the binary graph formalism cannot capture comprehensively their multi-scale organisation [5, 6]. For instance, we could not construct a network representing protein complexes and the ways they relate to each other if we only used binary relationships between proteins, as many of the proteins could not function in isolation outside the complex.

To overcome this limitation, we model an intra-cellular molecular system using a generalisation of graphs, hypergraphs [1, 5, 6]. In this new framework, representing a network of protein complexes comes naturally with each hyperedge corresponding to a set of proteins forming a specific complex, hence representing effectively protein complexes as illustrated on Figure 1. To uncover the functional information contained in the hypergraph-based models, it is necessary to develop tools to analyse the topology of hypergraphs.

2 Methods

In this project, we investigate hypernetworks that model multi-scale organisation of interacting molecules using an extension of graphlets to hypergraphlets. Jose Lugo-Martinez *et al.* [9] introduced a version of hypegraphlets, but the definition they used is not fit for relating the topology with biological function, since they relied on a non-standard definition of sub-hypergraph. Here, we correct for this and introduce hypergraphlets as small, connected, non-isomorphic, induced sub-hypergraphs of a real-world hypernetwork. We effectively use them to capture the local wiring patterns around a node in a hypergraph. Within a hypergraphlet, we also define *automorphism orbits* analogously to those in graphlets; informally, they can be viewed as nodes that can be swapped without changing the topology of the hypergraphlet. Considering all 1-node to 4-node hypergraphlets, there are a total of 472 different orbits. Figure 2 gives a representation of the 10 orbits occurring within the 2- to 3-nodes hypergraphlets.

We use hypergraphlets to extend graphlet-based statistics and analytics framework to mine several hypergraphs that model protein-protein interactions, protein complexes, biological pathways, drug-target data, and gene-disease data. In this study, first we use a clustering and enrichment analysis to investigate if similarly wired proteins in a hypernetwork have similar biological roles. Then we use a Canonical Correlation Analysis (CCA) to test if specific biological functions are performed by proteins having specific wiring in a hypergraph.

We demonstrate a relationship between the topology of these hypergraph representations of the multi-scale molecular organization and biological function unveiled by our new notion of hypergraphlets and tools that utilize them to mine these rich, complex data.

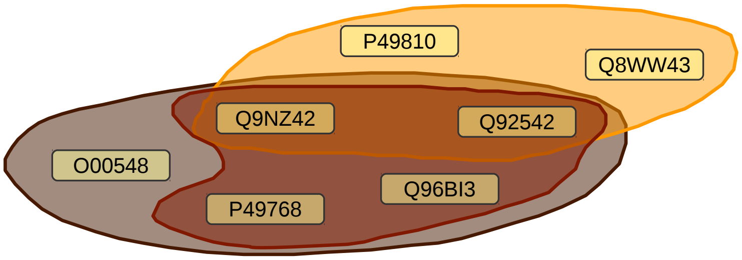


Fig. 1. An illustration of an hypergraph for three protein complexes that share proteins (identified by their UniProt IDs). The data was collected from CORUM database [7]. It represents complexes corresponding to sub-units of gamma secretase.

References

1. Berge, C. and Minieka, E., 1973. Graphs and hypergraphs. North-Holland Publishing Company Amsterdam.
2. Pržulj, N., Corneil, D.G. and Jurisica, I., 2004. Modeling interactome: scale-free or geometric?. Bioinformatics, 20(18), pp.3508-3515.
3. Pržulj, N., 2007. Biological network comparison using graphlet degree distribution. Bioinformatics, 23(2), pp.e177-e183.
4. Milenković, T. and Pržulj, N., 2008. Uncovering biological network function via graphlet degree signatures. Cancer Informatics, 6, p.257.

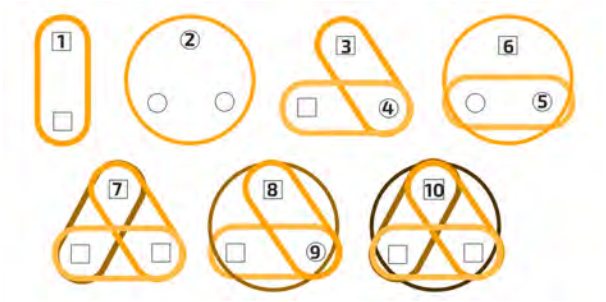


Fig. 2. One to three node hypergraphlets with the first eleven orbits. In a specific hypergraphlet, vertices represented with the same symbol (circle or square) belong to the same orbit.

5. Lacroix, V., Cottret, L., Thébault, P. and Sagot, M.F., 2008. An introduction to metabolic networks and their structural analysis. IEEE/ACM transactions on computational biology and bioinformatics, 5(4), pp.594-617.
6. Klamt, S., Haus, U.U. and Theis, F., 2009. Hypergraphs and cellular networks. PLoS Computational Biology, 5(5), p.e1000385.
7. Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. and Mewes, H.W., 2009. CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic acids research, 38(suppl_1), pp.D497-D501.
8. Yaveroğlu, Ö.N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A. and Pržulj, N., 2014. Revealing the hidden language of complex networks. Scientific Reports, 4.
9. Lugo-Martinez, J. and Radivojac, P., 2017. Classification in biological networks with hypergraphlet kernels. arXiv preprint arXiv:1703.04823.

Data-Fusion for Cancer Patient Stratification and Personalised Treatment

Vladimir Gligorijević[1], Noël Malod-Dognin[2], and Nataša Pržulj[2]

[1] Simons Center for Computational Biology, Flatiron Institute, New York 10010, USA
[2] Department of Computer Science, University College London, London WC1E 6BT, UK
E-mail: natasa@cs.ucl.ac.uk

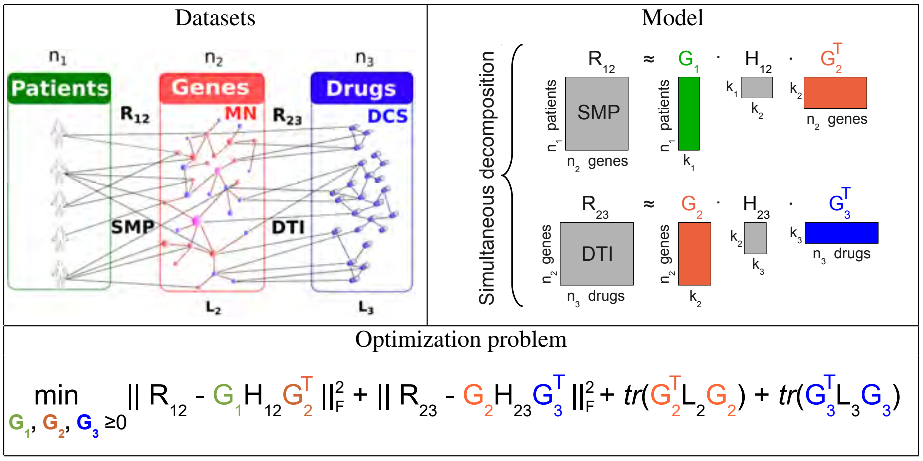
1 Background

Cancer is a leading cause of morbidity worldwide. It is a complex genetic disease in which the genomes of normal cells accumulate somatic mutations and other alterations that are eventually perturbing vital cellular functions. Recent advances in DNA sequencing technologies have enabled identification of somatic mutations across tumor genomes and exomes of individual patients [1]. These somatic mutations provide a new and rich source of data for addressing many challenges in cancer research, such as identifying driver genes (i.e., genes whose mutations lead progression of oncogenesis), stratifying patients into biologically meaningful classes with different clinical outcomes and creating new opportunities for development of successful personalized treatment strategies [2]. Cancer is also a highly heterogeneous disease with large genetic diversity even between tumors of the same cancer type. Namely, two clinically identical tumors rarely have a large set of common mutated genes. Moreover, very few genes are frequently mutated across tumor samples. This makes the use of somatic mutations for identification of driver genes, as well as for patient stratification into subtypes, much harder [3, 4, 1]. However, despite this genetic diversity between tumor samples, the perturbed pathways are often similar [1]. Therefore, integration of somatic mutations with other genomic data, such as with molecular networks that contain pathways, is a promising direction for addressing these problems.

2 Contributions

We present a versatile patient-specific data integration (fusion) methodology [5] capable of: 1) uncovering patient subgroups (stratification) with prognostic survival outcome, 2) predicting novel driver genes and 3) repurposing drugs, i.e., predicting new candidate drugs for targeting mutated gene products in individual patients and that can be used in treatment of identified patient subgroups. To our knowledge, this is the first method that can address all three challenges simultaneously.

Our methodology, is based on the Non-negative Matrix Tri-Factorization (NMTF) technique, initially proposed for dimensionality reduction and co-clustering problems in machine learning [6]. It approximates (factorises) a high-dimensional, $n_1 \times n_2$ data matrix R_{12}, representing relations between n_1 elements from type 1 and n_2 elements from type 2, as a product of three non-negative, low-dimensional matrices: $R \simeq G_1 H_{12} G_2^T$, where G_1 is the cluster indicator matrix for the first type (grouping the n_1 elements of type 1 into $k_1 \ll n_1$ clusters), G_2 is the cluster indicator matrix for the second type (grouping the n_2 elements of type 2 into $k_2 \ll n_2$ clusters), and H_{12} is the compressed representation of R_{12} that relates the n_1 clusters of type 1 to the n_2 clusters of type 2. The clustering interpretation of low-dimensional matrices and their previously established relatedness to the k-means clustering has enabled the use of NMTF in co-clustering problems [7]. Recently, there has been a significant development in the use of NMTF in data fusion because of its ability to extend to any number of interrelated data types by *simultaneously* decomposing their relation matrices. This has provided us with a valuable framework for fusion (integration) of any number and type of interrelated heterogeneous datasets [8, 9]. NMTF has demonstrated a great potential in addressing various biological problems, such as disease association prediction [8], disease gene discovery [10], protein-protein interaction prediction [11] and gene function prediction [12].



$$\min_{G_1, G_2, G_3 \geq 0} \| R_{12} - G_1 H_{12} G_2^T \|_F^2 + \| R_{23} - G_2 H_{23} G_3^T \|_F^2 + tr(G_2^T L_2 G_2) + tr(G_3^T L_3 G_3)$$

Fig. 1. Top left: The relationships between the integrated datasets. Somatic mutation profiles (SMP), encoded into matrix R_{12}, relate the patients to their mutated genes. The molecular interaction network (MN), encoded by its Laplacian matrix L_2, relate genes that interact together. Drug-target interactions (DTI), encoded into matrix R_{23}, relate drugs to the genes that they are targeting. Finally, drug chemical similarity network (DCS), encoded by its laplacian matrix L_3, relate chemically similar drugs. **Top right: Our data fusion model.** To integrate all data, we simultaneously decompose the somatic mutation profiles (matrix R_{12}) and the drug target interactions (matrix) into the product of lower dimensional matrix factors. The key point is in sharing the same matrix factor G_2 (the cluster indicator matrix of genes) across the decomposition, which allows learning from all datasets. **Bottom: Our optimization problem.** To simultaneously decompose R_{12} and R_{23}, we optimize the presented objective function using an iterative solver based on multiplicative update rules.

In our framework, which is illustrated in Figure 1, we use NMTF to integrate somatic mutation profile (SMP) data of 353 serous ovarian cancer patients from TCGA

[3] with molecular networks (MNs) from BioGRID[13] and KEGG [14], drug-target interaction (DTI) and drug chemical similarity (DCS) data from DrugBank [15]. We perform consensus clustering by using NMTF to simultaneously cluster patients, genes and drugs based on the evidence from *all* datasets. First, from the cluster indicator matrix of patients, G_1, we stratify patients into three groups. We observe significant difference in survival outcomes between these groups, as well as a good agreement with other clinical data. Second, from the cluster indicator matrix of genes, G_2, we identify clusters enriched in known driver mutations; we postulate genes strongly related to known driver genes in these clusters as potential drivers genes, i.e., genes responsible for ovarian cancer progression. Our predicted driver genes have good agreement with the literature. Finally, we use the matrix completion property of NMTF to predict new drug-target relations and to identify new drug candidates that could be used for repurposing and treatment of identified ovarian cancer patient groups. Furthermore, we evaluate the influence of all combinations of datasets onto the accuracy of drug-target predictions by performing a 5-fold cross validation. We show that the highest accuracy is achieved when all datasets are taken into account, proving the utility of integrating all considered datasets.

References

1. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Kinzler, K.W.: Cancer genome landscapes. Science 339(6127), 1546–1558 (2013)
2. Rubio-Perez, C., Tamborero, D., Schroeder, M.P., Antolín, A.A., Deu-Pons, J., Perez-Llamas, C., Mestres, J., Gonzalez-Perez, A., Lopez-Bigas, N.: In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer Cell 27(3), 382–396 (2015)
3. Cancer Genome Atlas Research Network: Integrated genomic analyses of ovarian carcinoma. Nature 474(7353), 609–615 (2011)
4. Hofree, M., Shen, J.P., Carter, H., Gross, A., Ideker, T.: Network-based stratification of tumor mutations. Nature Methods 10(11), 1108–1115 (2013)
5. Gligorijević, V., Malod-Dognin, N., Pržulj, N.: Patient-specific data fusion for cancer stratification and personalized treatment. In: Pacific Symposium on Biocomputing. p. 321332. World Scientific (2016)
6. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering. In: KDD ’06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data mining. pp. 126–135. ACM (2006)
7. Ding, C., He, X., Simon, H.D.: On the equivalence of nonnegative matrix factorization and spectral clustering. In: SDM. vol. 5, pp. 606–610. Proc. SIAM Data Mining Conf (2005)
8. Žitnik, M., Janjić, V., Larminie, C., Zupan, B., Pržulj, N.: Discovering disease-disease associations by fusing systems-level molecular data. Scientific Reports 3 (2013)
9. Žitnik, M., Župan, B.: Data fusion by matrix factorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on 37(99), 41–53 (2015)
10. Hwang, T., Atluri, G., Xie, M., Dey, S., Hong, C., Kumar, V., Kuang, R.: Co-clustering phenome–genome for phenotype classification and disease gene discovery. Nucleic Acids Research 40(19), e146–e146 (2012)
11. Wang, H., Huang, H., Ding, C., Nie, F.: Predicting protein–protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. Journal of Computational Biology 20(4), 344–358 (2013)

12. Gligorijević, V., Janjić, V., Pržulj, N.: Integration of molecular network data reconstruct gene ontology. Bioinformatics 30(17), i594–i600 (2014)
13. Chatr-aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M.S., Dolinski, K., Tyers, M.: The biogrid interaction database: 2015 update. Nucleic Acids Research 43(D1), D470–D478 (2015)
14. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28(1), 27–30 (2000)
15. Wishart, D.S., Knox, C., Guo, A.C., et al.: Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Research 36(suppl 1), D901–D906 (2008)

Understanding resilience in animal networks: the case of the macaque's social style

Ivan Puga-Gonzalez[a*], Sebastian Sosa[b*], and Cedric Sueur[a]

[a]Université de Strasbourg, CNRS, IPHC UMR 7178, Strasbourg, France.
ivanpuga@gmail.com cedric.sueur@iphc.cnrs.fr
[b]School of Sociology and Anthropology. Sun Yat-sen University, Guangzhou, 510275, China.
s.sosa@live.fr
[*]Both authors contributed equally in this work

1. Introduction

Group living animals rely on the transfer of information for an optimal exploitation of their habitat. Thus, the network structure of these societies should permit fast access to information to all individuals. Network's structures, however, may differ between societies or species, and these differences may affect information transmission and/or network resilience. For instance, in macaque societies, a taxon comprising ~20 species that live in multi-male/multi-female groups, network structure seems to differ according to dominance style: despotic or egalitarian. In despotic species, social networks are more modular, centralized, and less dense than networks of egalitarian species [1]. High modularity and centralization may produce structures similar to scale-free networks, and thus increase the risk of network disruption in case central individuals disappear. Conversely, in egalitarian societies, the low modularity and centralization of networks may produce more resilient networks. In support of this, two separate studies, one on Barbary macaques and other in chimpanzees (an egalitarian and despotic species respectively) showed that the network of Barbary macaques was resilient even after deleting 20% of the most central individuals; whereas that of chimpanzees was disrupted just after deleting the most centralized individual [2-3]. No study, however, has investigated whether there are systematic differences in the resilience of networks of despotic and egalitarian macaques.

The goal of the present study thus is to compare network resilience in despotic and egalitarian macaques. To do so, we make use of the individual-based model GrooFiWorld (Grooming and Fighting). We used this model because it reproduces network's features as well as behavioral patterns like those described in despotic and egalitarian macaques [1, 4-5]; and thus it makes it a perfect candidate to study and compare network resilience of despotic and egalitarian macaques while controlling for several confounding factors that cannot be controlled in nature (e.g. group composition, number of interactions).

2. Methodology

The model

GrooFiWorld is a spatially-explicit model in which individuals move and interact according to simple rules of thumb [4-5]. On encountering a partner, they can either interact in a negative (fight) or positive (groom) way. They fight if the risk of losing is low, otherwise they may groom its partner. These simple rules are enough to generate behavioral and network patterns similar to those of macaques.

Further, when intensity of aggression changes from high to low, the behavioral and network patterns also change, from those resembling despotic societies to those resembling egalitarian [1,4-5]. Using this model, we ran simulations at high and low intensity of aggression (to recreate ‘despotic’ and ‘egalitarian’ societies) and five different group sizes (10, 20, 30, 40, and 50). We ran 100 replicates per combination of parameters. A total of 600 grooming networks were built from the data collected (6000 interactions). Due to space limitations, here we only show the results for groups of 30 individuals; nevertheless, results remain qualitatively the same for all other group sizes.

Deletion simulations to compare resilience properties

To study network resilience, we deleted 20% of the most central nodes (those with the highest degree) in the network (target deletion) and compared the resulting network structure with that obtained when deleting nodes at random (random deletion). After each deletion, we measured the following global network metrics: 1) diameter, 2) clustering coefficient, 3) modularity, and 4) global efficiency. To compare the differences between target and random deletions, we ran Generalized Linear Mixed Models (GLMM) with networks as random factors to deal with repeated measures on network global metrics over the successive node deletions. The dependent variable was the global metric and the predictor variable the type of deletion. Similarly, to compare resilience of the networks according to dominance style, we also ran GLMM with networks as random factor. The dependent variable was the change in the global metric after target deletion, and the predictor variable was dominance style: despotic or egalitarian.

3. Results

GLMM showed significant differences between target and random deletions for all network metrics and for both despotic and egalitarian societies (Tables I, II). In target deletions, the diameter and modularity of the network increase whereas clustering coefficient and global efficiency decrease more rapidly than in random deletions (Table I).

		DESPOTIC			EGALITARIAN		
Metric	Factor	Value (SE)	t-value	p-value	Value (SE)	t-value	p-value
Diameter	Intercept	3.54 (0.06)	56.52	<0.001	2.15 (0.02)	99.27	<0.001
	Target vs Random	0.13 (0.01)	11.77	<0.001	0.02 (0.00)	6.29	<0.001
CC	Intercept	0.63 (0.00)	166.47	<0.001	0.71 (0.00)	220.40	<0.001
	Target vs Random	-0.02 (0.00)	-28.07	<0.001	-0.01 (0.00)	-16.17	<0.001
Mod	Intercept	0.20 (0.01)	31.08	<0.001	0.18 (0.01)	26.67	<0.001
	Target vs Random	0.04 (0.00)	27.93	<0.001	0.01 (0.00)	6.88	<0.001
GE	Intercept	0.94 (0.01)	133.26	<0.001	1.07 (0.01)	196.28	<0.001
	Target vs Random	-0.04 (0.00)	-24.20	<0.001	-0.01 (0.00)	-11.31	<0.001

Table I. *GLMM between target and random deletions in despotic and egalitarian artificial societies (n=30). CC = clustering coefficient; Mod= modularity; GE= global efficiency.*

Further, in despotic societies, the change in the value of the network metrics was greater than in egalitarian societies (Table II). However, we did not observe network disruption, neither in despotic nor in egalitarian societies.

4. Discussion

In the present study, we investigated whether network resilience differed between artificial despotic and egalitarian macaque societies. Our results showed that in both types of societies, deleting the most central nodes had a more deleterious effect than when nodes were deleted at random; suggesting then, that central individuals aid in the maintenance of group cohesiveness and fast information transmission. When comparing networks of despotic and egalitarian, we found that after target deletions networks of despotic societies became less efficient than those of egalitarian; however, in no case network disruption was observed. This suggests that despite differences in network structure, networks of egalitarian and despotic societies may be adapted to be resilient. Our results, however, should be taken with caution since they await empirical confirmation. If confirmed, they indicate that network's properties making it resilient may be favored by natural selection, and that a trade-off may exists between centralized societies with fast information transmission and decentralized societies with low information transmission but great cohesiveness. Whether this may also constraint the individuals' behavioral patterns underlying network structure deserves further investigation.

Metric	Factor	Value (SE)	t value	P-value
Diameter	Intercept	0.55 (0.03)	17.08	<0.001
	Ega vs Des	-0.45 (0.05)	-9.97	<0.001
CC	Intercept	-0.04 (0.00)	-21.86	<0.001
	Ega vs Des	0.03 (0.00)	10.66	<0.001
Mod	Intercept	0.09 (0.00)	20.60	<0.001
	Ega vs Des	-0.07 (0.00)	-10.94	<0.001
GE	Intercept	-0.12 (0.00)	-64.99	<0.001
	Ega vs Des	0.09 (0.00)	35.03	<0.001

Table II. *GLMM between despotic and egalitarian artificial societies (n=30). CC = clustering coefficient; Mod= modularity; GE= global efficiency; Ega= egalitarian; Des= despotic.*

References

1.Sueur C., Petit O., De Marco A., Jacobs A., Watanabe K., Thierry B. A comparative network analysis of social style in macaques. Anim Behav 82(4), 845-852 (2011)

2. Kanngiesser P, Sueur C, Riedl K, Grossmann J, Call J (2011) Grooming network cohesion and the role of individuals in a captive chimpanzee group. Am J Primatol 73:758-767

3. Sosa S (2014) Structural Architecture of the Social Network of a Non-Human Primate *(Macaca sylvanus):* A Study of Its Topology in La Foret des Singes, Rocamadour. Folia Primatol 85:154-163

4.Puga-Gonzalez I., Sueur C. Emergence of complex social networks from spatial structure and rules of thumb: a modelling approach. Ecol Comp 31, 189-200 (2017)

5.Puga-Gonzalez I., Hildenbrandt H., Hemelrijk C.K. Emergent patterns of social affiliation in primates, a model. PLoS Comput Biol 5(12), e1000630 (2009)

Gene interaction network to prioritize gene selection using Markov Random Fields model

Rémi Souriau^{1,2}, Jaakko Nevalainen^{3}, Guillaume Pinna^{4}, Florent Chatelain^{2}, and Laurent Guyon^{1*}

^{1}Univ. Grenoble Alpes, CEA, INSERM, BIG, BCI UMR_S 1036, F-38000 Grenoble, France
^{2}GIPSA-Lab, Department of Images and Signal, F-38402 Grenoble, France
^{3}School of Health Sciences, University of Tampere, Tampere, Finland
^{4}Plateforme ARN interférence, IBITEC-S, CEA, F-91191 Gif-sur-Yvette, France
*laurent.guyon@cea.fr

1 Introduction

Following the sequencing of the human genome in 2003, we entered in the genomic era. Thanks to the automation of biological processes, biologists are now able to generate lists of p-value associated genes containing thousands of genes. These p-values allow one to reject the null hypotheses that genes are not related to a phenomenon of interest. For instance, the p-value is the probability that the corresponding gene product has such a low concentration in the cell or such a small activity if the null hypothesis was true. Extreme p-value supports the rejection of the null hypothesis. High throughput technologies used to generate these gene lists are often imperfect, leading to false detections. Here we will exemplify the methodology using an RNA interference (RNAi) screen. Such screens are called functional screening; each gene product is repressed or induced and the cell fate related to the biological process of interest is further quantified and transformed into a p-value. For example, if after gene repression, cells proliferate significantly more than the negative control case, a low p-value will be associated to this gene.

Here we propose to take into account known interactions between pairs of genes, organized in a network, to prioritize gene selection in RNAi functional screening data. In the network, nodes are genes and the links $e_{i,j}$ correspond to pairs of interacting genes i and j. We use undirected networks, either weighted ($e_{i,j}$ is a positive scalar) or not ($e_{i,j}$ equal 0 or 1). The aims of this network based prioritization are, among others, to reduce false discovery rate (FDR) and to facilitate biological pathway identification.

Our hypothesis is that interacting genes have higher chance to have similar p-values. In the dataset analyzed, the observed extreme values are two-sided. However, we perform a one-sided test, leading to enriched amounts of genes associated with low p-values (called negative hits) and with high close to 1 p-values (positive hits), as shown Fig. 1a. Our aim is then to select in the network positive and negative gene hits separately that are clustered together. We herein present and extend our work recently published [5].

2 Network segmentation with Markov Random Field energy minimization

For each gene, we aim at identifying a label, either 1 (positive hit), -1 (negative hit) or 0 (non-hit), based both on its own p-value and the p-values of its neighbor genes in the network. Similar to what was done in image segmentation with a simpler and regular network [3], we propose to use a Markov Random Field (MRF) model, considering each node depends only on its first neighbors in the graph. With three labels, it corresponds to the Potts model, being an extension of the 2-labels Ising model as introduced in statistical physics.

Let X be the vector of -1/0/1 labels to be inferred and Z the vector of observed p-values. We define the following energy, sum of a potential energy E_{pot} depending solely on p-values and an interaction energy E_{int} depending on the network only, to be minimized:

$$E = E_{pot} + E_{int} = \sum_i \sum_k -\log(f_k(z_i)).I_d(x_i = k) + \beta \sum_i \sum_j e_{i,j} I_d(x_i \neq x_j) \tag{1}$$

The first term is the sum of an individual energy for each gene i. If the gene is a non-hit (label $x_i = 0$), its p-value follows a uniform f_0 probability density and the corresponding energy is zero. If gene i is a hit ($x_i = k = \pm 1$), its potential energy is $-\log(f_k(p_i))$. Here f_k is the density function of one of the two alternative hypothesis, for which we chose a *beta* law with parameters (1,a) or (a,1): $f_{-1}(z) = az^{a-1}$ or $f_1(z) = a(1-z)^{a-1}$ with $0 < a \leq 1$. The case $a = 1$ corresponds to the uniform distribution, and decreasing a values lead to more peaked distributions around zero (Fig. 1a). Interestingly, this parameter a can be related with False Discovery Rate (FDR) for the hit detection when the prior information given by the network is not taken into account ($\beta = 0$). This yields a simple way to either calibrate this parameter or to interpret the detection results for a given value of a. If $p_i < p_i^*$, where p_i^* is the p-value corresponding to $f_{-1}(p_i^*) = 1$, the potential energy minimization will favor the negative hit label ($x_i = -1$). The second term, the network *a priori*, is parametrized by β. If two neighbor genes i and j have different labels, the energy function provides a penalization (energy > 0) equal to $2\beta e_{i,j}$, where $e_{i,j}$ is the weighted score of the link in the network. We then define $\hat{x}$ as the vector minimizing this energy: $\hat{x} = \text{argmin}_x\, E(x)$.

Classically, the energy can be minimized using Gibbs sampling or graph cut procedures [2]. We built an R wrap-up from the `maxflow` C++ library for the graph cut procedure. This library is developed by the authors of the BK algorithm [1].

3 Results

We showed previously that this procedure performs the best among other published methods using independent simulated data, and that in an RNAi screen relevant biological processes were found enriched among the list of hit genes [5]. Increasing β leads to more clustered genes with the same label until $\beta = \beta*$ where the same label is assigned to all genes. This 'dominant' label is the non-hit ($k = 0$) as we expect more genes do not participate to the biological process of interest. A score was then defined as the amount of tested β values for which the gene was a hit further multiplied by the degree of the gene. For the sake of visualization, Fig. 1b shows a smaller 500 nodes network overlaid with observed p-values obtained in a non-coding gene RNAi screen. Fig. 1c shows the same number of hit genes with the top MRF score for each label, which are clearly more clustered but biased toward high degree nodes. We are currently normalizing E_{int} by $d_i = \sum_i e_{i,j}$ to overcome this bias. We will also discuss various strategies to handle the unknown parameter β including Bayesian approaches to estimate β and marginalization of β as proposed by [4].

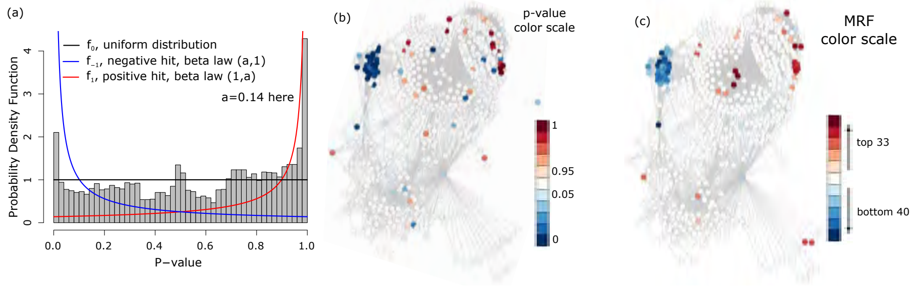


Fig. 1. (*a*) Corrected median P-value distribution of the gene data as in [5] superposed with density function chosen for each label (negative hits f_{-1}, positive hits f_1 and no hit f_0). (*b*) and (*c*) 500-gene binary network in which functionally related genes are connected. (*b*) Observed p-values, with a blueish (resp. reddish) color scale for genes with p-value below 0.05 (resp. above 0.95), likely to be negative (resp. positive) hits. (*c*) Corresponding MRF score hits (same number of nodes are colored), false colors for positive and negative hit labels.

Acknowledgments FC, RS and LG received support from the AGIR grant ABC_ARNi from Univ. Grenoble Alpes. LG thanks Sean Robinson for interesting remarks.

References

1. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 1124–1137 (September 2004)
2. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence 26(2), 147–159 (Feb 2004)
3. Li, S.Z.: Markov Random Field Modeling in Image Analysis. No. 3 in Advances in Computer Vision and Pattern Recognition, Springer-Verlag London (2009)
4. Pereyra, M., McLaughlin, S.: Fast unsupervised bayesian image segmentation with adaptive spatial regularisation. IEEE Transactions on Image Processing 26, 2577–2587 (2017)
5. Robinson, S., Nevalainen, J., Pinna, G., Campalans, A., Radicella, J.P., Guyon, L.: Incorporating interaction networks into the determination of functionally related hit genes in genomic experiments with markov random fields. Bioinformatics 33(14), i170–i179 (2017)

Perturbation of amino acid networks: A statistical study of the defects introduced in proteins by mutations

Rodrigo Dorantes-Gilardi[1,2], Laurent Vuillon[2], and Claire Lesieur[1,3]

[1] IXXI-ENS Lyon, Lyon, France
[2] Laboratoire de Mathématiques, LAMA UMR 5127,
Université Savoie Mont Blanc,CNRS, Le Bourget du Lac, France
[3] Laboratoire AMPERE, UMR5005,
CNRS-UCBL-INSA-ECL, Villeurbanne, France

1 Introduction

Proteins are molecules that efficiently carry out biological activities. These activities rely on the protein structure, which has evolved for a specific functional role [1]. The structure of a protein is based on the chemical interactions between the atoms of the amino acids, building blocks of proteins. Unfortunately, the exact atomic interactions within proteins are intractable due to the complexity of the system. For about fifteen years, a successful alternative has been to model protein structures as a network of amino acids in interaction (spatial network) [3]. An amino acid network (ANN) represents amino acids as nodes where each link connects two nodes if the respecting amino acids have at least one pair of atoms at distance less than a given cutoff. Amino acid networks (AANs) have been used to investigate topics which include protein-protein interactions, communication within and between proteins, and protein-folding [3].

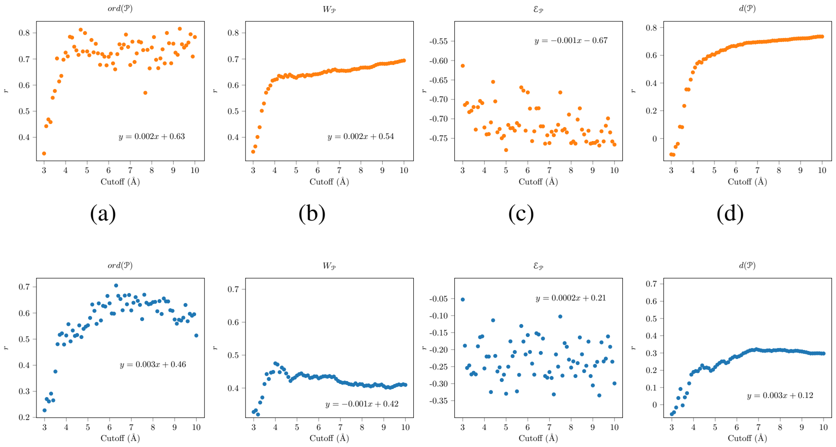
So-called functionally sensitive positions (FSPs) tend to undermine the functional activity of the protein when mutated by other amino acids, often to adapt an alternative function [2]. Here, we aim at investigating the particular *local structures* of FSPs relative to robust positions.

The structural changes associated with mutation cover orders of magnitude from ångström (10^{-10}m) to nanometer (10^{-9}m), yet commonly AANs are built using a single cutoff corresponding to chemical distances to have atomic interactions. Since structural changes extend on multiple lenght-scale, the choice of a unique chemical cutoff is somehow arbitrary and may miss collective behaviour which implies amino acid communication at distance above atomic interactions. We have investigated the use of many cutoffs to analyze the structural impact of mutations, and found it to be good practice and offer qualitative and quantitative information on the mutation effects.

2 Methods and Results

The case study is the third PDZ domain from the PSD-95 protein, X-ray-obtained 3D structure available with code 1BE9. AAN is constructed from the calculation of all atom-atom distances in the protein. Each atomic pair is connected by a link if the atoms are at distance less than a given cutoff. Subsequently, groups of nodes (atoms) belonging to one amino acid are collapsed into a single node. The links between two nodes of the resulting multi-graph are then collapsed into a single edge with a weight equal to the number of links collapsed, finally obtaining the AAN. The properties were computed for 71 distance cutoffs.

We mutated *In Silico* each position of the protein by the other 19 amino acids obtaining a database of $83 \times 19 = 1577$ mutated 3D structures. For each mutated structure, we computed a perturbation network $\mathcal{P}$ by comparing its ANN with the original ANN: $\mathcal{P}$ contains the set of nodes incident to a link with different weight in the two networks. Each link in $\mathcal{P}$ either has different weight in the mutated and the wild type networks, or it exists in only one of the networks. The functional change values for the same mutational database were taken from the experimental work by McLaughlin et al [2].



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 1: Structural and functional relation between the protein and $\mathcal{P}$. The correlation between buriedness of a position and the order, total weight, Euclidian distance, and degree of its perturbation network $\mathcal{P}$ is shown in a, b, c, and d, respectively. The second row of figures e, f, g, and h shows the correlation of the same properties with functional change.

We considered the following network properties of $\mathcal{P}$: the order of $\mathcal{P}$, namely its number of nodes, noted $ord(\mathcal{P})$; the weight of $\mathcal{P}$, equal to the sum of the weights of its links, noted $W_{\mathcal{P}}$; and the largest Euclidian distance in the structure between the mutated node and any other node of $\mathcal{P}$, noted $\mathcal{E}_{\mathcal{P}}$. Finally, we considered the degree of each node corresponding to the mutated amino acid, noted $d(\mathcal{P})$. For each network property, the average value of the position over the 19 other possible amino acid type is calculated. To have a clear indication of the relevance of the network properties to the protein structure, we computed their correlation to the so-called buriedness of the protein, where the buriedness of each position is defined by its Euclidian distance to the protein surface.

The results show a clear correlation between all four $ord(\mathcal{P})$, $\mathcal{E}_{\mathcal{P}}$, $d(\mathcal{P})$, and $W_{\mathcal{P}}$with the protein buriedness (Fig. 1 a, b, c, and d). The position spatial location in the protein measured by its buriedness is a clear indicator of a its value in terms $\mathcal{E}_{\mathcal{P}}$, $ord(\mathcal{P})$, $d(\mathcal{P})$, and $W_{\mathcal{P}}$. Interestingly, the correlation between the buriedness of a position and $\mathcal{E}_{\mathcal{P}}$ is negative, implying that mutations at surface positions "perturb" more distance neighbors than buried positions. The correlation between buriedness and $d(\mathcal{P})$ has a more straightforward interpretation as the weighted degree of a position, together with its potential change in the mutant AAN, is larger for fully surrounded positions. Finally, strong correlation between buriedness and both $ord(\mathcal{P})$ and $W_{\mathcal{P}}$ indicates a larger $\mathcal{P}$ for buried positions in terms of both nodes and links.

On the other hand, $ord(\mathcal{P})$ is the only property of the network showing a good correlation with experimentally measured functional change (Fig. 1 e). The number of nodes of $\mathcal{P}$ explains up to 70% of the functional damage, Since the three other network properties do not correlate with functional damage, it suggest that $ord(\mathcal{P})$effect is not simply due to the buriedness of the position (Fig. 1 f, g, and h).

Conclusion Buried nodes are captured by all the perturbation network properties considered, independently from the cutoff chosen. This shows the stability of the measures to model the protein structure. The strong correlation between the number of nodes in the perturbation network and the functional damage by mutations, together with the lack of correlation between the degree of the node mutated and the functional damage, do not seem to be due to structural damages at the position site but to the extent of structural changes. What counts is not the connectivity of the site of mutation (single fragile node), but its influence over the network (collective fragility). Furthermore, it is possible to assess functional fragility using the network properties of the perturbation network resulting from the comparison of the amino acid and wild type networks, as long as we use multiple cutoffs to guarantee the stability of the properties. Finally, this approach allows to discriminate structural fragility in two categories relative to the perturbation network: number of nodes perturbed by the mutation, independently from the degree of the node, and number of weights affected overall. Now we will look at the features of these two classes of fragile nodes together with additional parameters of the perturbation network to understand the mechanisms of structural responses underlying functional fragility.

References

1. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science 181(4096), 223–230 (1973)
2. McLaughlin Jr, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., Ranganathan, R.: The spatial architecture of protein function and adaptation. Nature 491(7422), 138 (2012)
3. Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G., Shen, B.: The construction of an amino acid network for understanding protein structure and function. Amino acids 46(6), 1419–1439 (2014)

Covariate-Adjusted Binary Ising Model

Jai Woo Lee[1], Margaret R. Karagas[2] and Jiang Gui[1, 3]

1 Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH,
f001c2p@dartmouth.edu
2 Department of Epidemiology, Geisel School of Medicine, Lebanon, NH,
margaret.r.karagas@dartmouth.edu
3 Department of Biomedical Data Science, Geisel School of Medicine, Lebanon, NH,
jiang.gui@dartmouth.edu

1 Introduction

When we try to find associations of elements in the network, considering covariates specifically correlated with some subsets of elements in the network can help to define hidden but significant interactions among elements. In the real data, we can assume a few epidemiological factors such as dietary intakes as covariates, and using those covariates can define new relations among gut microbiota elements strongly correlated to specific dietary intake information. To implement this experiment, we utilize binary Ising model which conducts logistic regression with variable selection based on validation of models to define interactions between two elements in a network [1]. To obtain the optimal model, Elastic Net or Lasso penalized regression methods with validation methods such as AIC, AICc, or BIC are tested. Our developed method can be applied to generate the network of gut microbiota with relevant epidemiologic factors from the New Hampshire Birth Cohort Study. Our goal is to investigate computationally and biologically meaningful relations among targeted covariates and relevant elements in regression and also to develop effective adjustment methods.

2 Methods and Data

Let us consider one undirected network with **n** nodes and **k** covariate nodes of degree **d**. Each of **k** covariate nodes is strongly correlated with distinct groups of **d** nodes each in the network. Each of **n** nodes consists of **p** binary samples. To define the interactions among elements, the proposed binary Ising model considers the logistic regression model and to get optimal coefficients for this model, implements penalized logistic regression for binary data by minimizing negative log-likelihood with penalty [2]. For binary logistic regression, log-likelihood is defined as following:

$$\sum_{i=1}^{n+k} y_i X_i \beta - \log(1+\exp(X_i\beta)) \qquad (1)$$

where y_i is the i^{th} element in response variable, X_i is the i^{th} vector in input matrix, and β is the estimated coefficient vector. The optimal model is determined by the best resulting set of coefficients from the following three criteria. Here, df indicates the number of estimated coefficients or interactions by the logistic regression method:

With AIC,

$$argmin_\beta(-2 * \sum_{i=1}^{n+k} y_i X_i \beta - \log(1+\exp(X_i\beta)) + 2 * df) \qquad (2)$$

With AICc,

adfa, p. 1, 2011.

$$argmin_{\beta}(-2 * \sum_{i=1}^{n+k} y_i X_i \beta - \log(1+\exp(X_i\beta)) + 2 * df * \frac{df+1}{p-df-1}) \quad (3)$$

With BIC,

$$argmin_{\beta}(-2 * \sum_{i=1}^{n+k} y_i X_i \beta - \log(1+\exp(X_i\beta)) + \log(p) * df) \quad (4)$$

To include k covariate nodes always in regression, we use penalty factors to control penalized nodes. Below is the formula for penalty factors:

$$\lambda \sum_{j=1}^{n+k} v_j [(1-\alpha)\frac{1}{2}\beta_j^2 + \alpha|\beta_j|] \quad (5)$$

where v_j denotes the penalty factor for j^{th} element, α determines the selection of Lasso (α=1), Elastic Net (0< α <1), or Ridge (α=0) and β_j indicates the estimated coefficient for j^{th} element. We set $v_j = 0$ to make k covariates unpenalized and set v_j=1 to make n nodes penalized.

The binary Ising model implements logistic regression with model selection using a goodness-of-fit measure to define relevant relationships between elements indicating interactions in the same network [3]. Given a data matrix with sample size p and n nodes, we can formulate the overall formula,

$$\sum_{i=1}^{n+k}(y_i X_i \beta - \log(1+\exp(X_i\beta)) + \ \lambda \sum_{i=1}^{n+k} v_i [(1-\alpha)\frac{1}{2}\beta_i^2 + \alpha|\beta_i|]) \quad (6)$$

The binary Ising model assumes samples of each element are classified into -1 or 1; In this study, very small values in each sample of element are classified as 0 and all the other values are classified as 1. With the penalty factor non-penalizing covariates, we utilize Ising model to find hidden relations among elements relevant with the specific covariate in the network. After non-penalizing covariates, we examine interactions of sample of each element and samples of all other existing elements by considering logistic regression and binary Ising model.

The data of 1H NMR metabolomic profiling were collected from 386 subjects from the New Hampshire Birth Cohort and information on nutrition, environmental exposure, or maternal life style was available on participants. Including available gut microbiota data with relevant epidemiologic information from participants in simulation, we utilized

our method to resolve one important problem indicating whether given epidemiologic information interacting with each specific group of gut microbiota elements, our method can construct distinct subnetworks of gut microbiota elements interacting with similar epidemiologic factors.

3 Results

3-1) A Toy Example with Simulation Results: 400 samples for 15 element nodes and 3 covariate nodes each having 3 element node neighbors, estimated with Lasso and AICc

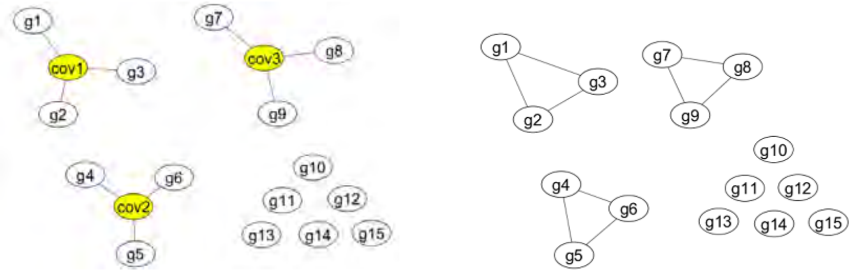


Fig. 1. The network on the left side includes 15 element nodes in total represented as g and 3 covariate nodes represented as cov; each covariate node is connected to a distinct group of 3 element nodes. It is expected that if we weaken or remove the interactions between element and covariates, the element nodes previously connected to covariates would form new interacting edges as shown on the right side. In the real data, g can be considered as gut microbiota nodes and cov can be considered as dietary information.

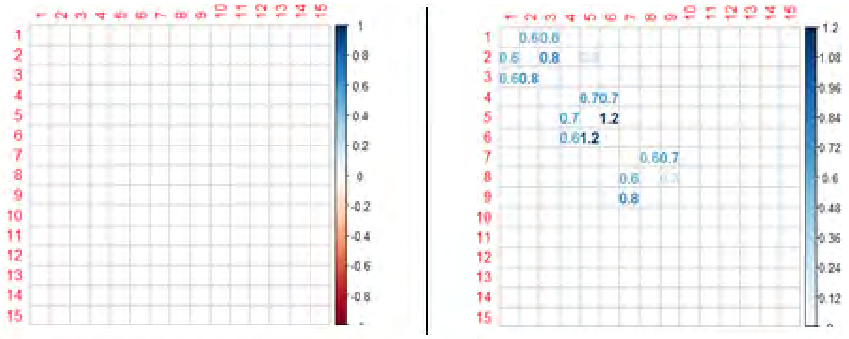


Fig. 2. The matrix on the left side shows that if we do not consider the effect of covariates in the network, no new associations among 15 element nodes would be detected. The matrix on the right side shows that internal interactions among element nodes 1-3, 4-6, or 7-9 each are detected if we use covariate adjustment.

3-2) Simulation Results with a Large Network: various samples for 200 element nodes and 5 covariate nodes each having 5 element node neighbors, estimated with Lasso and AICc

As shown in Fig. 1 and Fig. 2, covariate-adjustment helps to track new interactions significantly correlated with covariates but possibly hidden when we do not consider covariate-adjustment. In the large network with at least 200 element nodes, there are more parameters which we should consider. Fig. 3 shows us that choosing appropriate sample size is important to detect associations in the large network [4]. Plus, the number of element nodes correlated with covariates affects estimation of correct edges. The 200 element node network of sample size 5000 with covariates of degree 4 gives 0.99 balanced accuracy, one with covariates of degree 5 gives 0.91 balanced accuracy and one with covariates of degree 10 gives 0.5 balanced accuracy with AICc and Lasso models. So far, there are no other identified parameters significantly affecting balanced accuracy measurement.

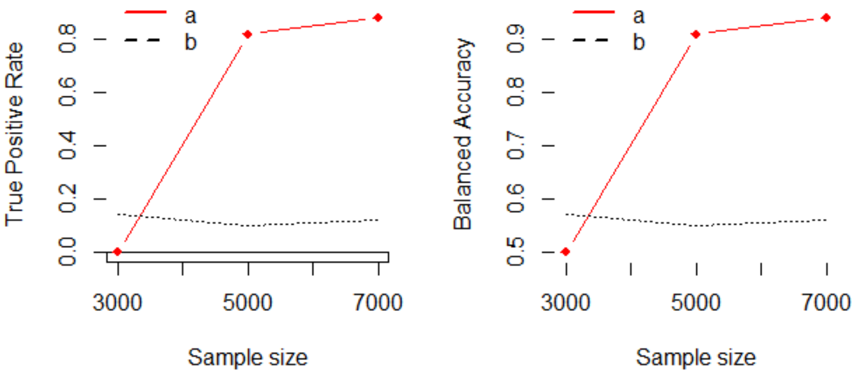


Fig. 3. For the measurement shown in this figure, we consider that if our method correctly selects edges among element nodes correlated with the same covariate node as represented in Fig. 1, correctly chosen edges are true positive cases. Also, the balanced accuracy, the average of true positive rate and true negative rate, was used to measure the estimation accuracy. The line graph on the left side indicates how sample size affects true positive rates and one on the right side indicates how sample size affects balanced accuracy. The red line 'a', indicates the method considering covariate adjustment method and the black line, 'b', indicates the method not considering covariate adjustment method. With sufficient sample size, implementing covariate-adjustment helps to find new significant interactions among element nodes

For future directions, we consider computational experiments including different parameters such as sample size, degree of nodes, the number of strongly influencing covariates for each node, or randomness in the network. Also, application of minimum spanning or matching algorithm which can extract necessary and significant interactions from resulting networks is studied. Finally, gut microbiota of more interest and essential dietary or any other epidemiologic information would be examined.

References

1. Li, X., Du, J. , Li, G., Fan, M.: Variable selection for covariate adjusted regression model. Journal of Systems Science and Complexity, Volume 27, Issue 6, 1227-1246 (2014)

2. Friedman, J., Hastie, T., Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent (2010)

3. van Borkulo, C., Borsboom, D., Epskamp, S., Blanken, T., Boschloo, L., Schoevers, R., Waldorp, L.: A new method for constructing networks from binary data. Scientific Reports 4, Article number: 5918 (2014)

4. Ravikumar, P., Wainwright, M., Lafferty, J.:High-dimensional Ising model selection using l1-regularized logistic regression. Annals of Statistics, Vol. 38, No. 3, 1287-1319 (2010)

A Network-Based Metapopulation Model to Simulate a Pulmonary Tuberculosis Infection

Michael Pitcher[1], Ruth Bowness[2], Simon Dobson[1], and Stephen Gillespie[2]

[1] School of Computer Science, University of St. Andrews, Jack Cole Building, North Haugh, St. Andrews, Fife, KY16 9SX
mjp22@st-andrews.ac.uk
simon.dobson@st-andrews.ac.uk
[2] School of Medicine, University of St. Andrews, North Haugh, St. Andrews, Fife, KY16 9TF
rec9@st-andrews.ac.uk
shg3@st-andrews.ac.uk

1 Introduction and background

Tuberculosis (TB) is an infectious disease that claims over 1 million lives globally [8]. A TB infection typically occurs in the lungs, with *Mycobacterium tuberculosis* (M.tb) bacilli being inhaled and lodging in the alveolar tissue, where they begin a long and complex interaction with the human immune system. The development of novel treatments for TB is hampered by the lack of an animal model that completely reflects the pathology that occurs in humans with the disease [5]. Computational models allow the creation of an artificial human environment in which to test hypotheses and explore the functions and environments of an infection.

The lung environment plays an important role in the progression of TB disease. Initial infection typically occurs in the lower, well-ventilated regions of the lung [7]. However, reactivation of a primary disease that has been previously contained by the body's immune system almost always occurs at the apices of the lung [3, 6]. It has been hypothesised [1] that the apical regions provide a preferable environment for the bacteria, with attributes such as a high oxygen tension, low immune activity and low lymphatic drainage providing the ideal requirements for bacterial growth. As such, bacteria that implant in the lower region of the lungs must disseminate, possibly via the lymphatics [2], to the apical regions in order to find the best situations to proliferate and spread to others.

The exact impact that this form of spatial heterogeneity and bacterial dissemination have on the progression of TB are poorly understood. In this work, we present a framework to create network-based metapopulations with spatial heterogeneity. This framework is then used to create a whole-organ model of the lung as the environment upon which to model TB dynamics.

2 Model

ComMeN (Compartmentalised Metapopulation Network) is a Python-based framework to allow for simple creation of network-structured metapopulations. A network is created, the nodes of which contain a subpopulation split into distinct compartments and both the nodes and edges may include spatial attributes, allowing for a spatially heterogeneous environment. A series of situation-specific events can be created which determine how the members in each patch interact with each other and how members translocate from one patch to another. The system uses a form of discrete-event simulation (originally detailed in [4]) to stochastically determine which events occur. Simulations run until a time limit is reached or no more events can possibly occur.

The framework is then extended to create PTBComMeN (Pulmonary TB ComMeN), which models the lung environment as a metapopulation and allows TB disease dynamics to be applied across it. The nodes of the network model regions of the lung tissue, each of which include spatial attributes such as oxygen availability and blood perfusion. All nodes within the lung network are joined with an edge, with edge weights signifying the possibility of translocation between them via direct transfer along the bronchi. A single node is included to model the lymphatic system and this is connected to all lung patches. The compartments in subpopulations of each node include the bacteria and a variety of immune cells, each split into various states depending on infection and activation statuses. The events included determine how compartments interact with one another (e.g. ingestion of bacteria by immune cells) in nodes and how translocation from one node to another occurs.

3 Methods

PTBComMeN allows for the simulation of TB dynamics across the whole lung environment, whereby we can focus on two important aspects: i) what effect does a supposedly preferential environment at the apical regions of the lung have on the development of a TB infection and ii) which means of dissemination are important to reaching this apical location. The network will be seeded with the native immune cells and a small number of bacteria, whose initial location is determined by the ventilation of each Lung patch. Simulations will then be run, with a focus on determining how critical dissemination into the lung apices is to establishing disease within the patient. We intend to validate these results against existing clinical data, such as x-rays, of TB patients. Future iterations of the model will integrate pharmacokinetic and pharmacodynamic data to model the effects of drug treatment of TB.

References

1. Balasubramanian, V., Wiegeshaus, E., Taylor, B., Smith, D.: Pathogenesis of tuberculosis: pathway to apical localization. Tubercle and Lung Disease 75(3), 168–178 (jun 1994), http://www.tuberculosisjournal.com/article/0962-8479(94)90002-7/
2. Behr, M.A., Waters, W.R.: Is tuberculosis a lymphatic disease with a pulmonary portal? The Lancet Infectious Diseases 14(3), 250–255 (2014), http://dx.doi.org/10.1016/S1473-3099(13)70253-6
3. Elkington, P.T., Friedland, J.S.: Permutations of time and place in tuberculosis. The Lancet Infectious Diseases 15(11), 1357–1360 (nov 2015), http://www.sciencedirect.com/science/article/pii/S1473309915001358

4. Gillespie, D.T.: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 22(4), 403–434 (1976), http://www.sciencedirect.com/science/article/pii/0021999176900413
5. Guirado, E., Schlesinger, L.S.: Modeling the Mycobacterium tuberculosis granuloma - the critical battlefield in host immunity and disease. Frontiers in immunology 4(April), 98 (2013), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3631743/
6. Hunter, R.L.: Tuberculosis as a three-act play: A new paradigm for the pathogenesis of pulmonary tuberculosis. Tuberculosis 97, 8–17 (2016), http://dx.doi.org/10.1016/j.tube.2015.11.010
7. Van Dyck, P., Vanhoenacker, F.M., Van den Brande, P., De Schepper, A.M.: Imaging of pulmonary tuberculosis. European Radiology 13(8), 1771–1785 (2003)
8. World Health Organization: Global Tuberculosis Report. Tech. rep., World Health Organization (2016), http://www.who.int/tb/publications/global_report/en/

A network analysis of the incidence pattern of microcephaly in the context of Zika Virus Infection

Authors: Myriam Patricia Cifuentes[1,2*], Clara Mercedes Suarez[3], Ricardo Cifuentes[4], Nathan Doogan[2], Noel Malod-Dognin[5], Sam Windels[5], Jose Fernando Valderrama[6], Darryl B. Hood[2], Natasa Przulj[5]

Affiliations:
[1]Department of Mathematics, College of Sciences, Universidad Antonio Nariño, Colombia.
[2]Division of Environmental Health Sciences, College of Public Health, The Ohio State University, U.S.A.
[3]Maestria en Salud Pública, Universidad Santo Tomas.
[4]Universidad Militar Nueva Granada.
[5]Department of Computer Science, University College London, UK.
[6]Sub-directorate of Transmissible Diseases, Ministerio de Salud, República de Colombia.

* mpcifuentes@uan.edu.co, mpcifuentesg@unal.edu.co

1 Introduction:

During the 2015-2016 epidemic of the Zika Virus (ZIKV) in Latin America, the dissimilar geographical distribution of the associated microcephaly condition raised questions about the virus being the sole cause of the birth defect [1]. In Brazil, the most affected country, synergies between socio-economic and environmental factors, such as education level and the use of agro-toxics, were suggested as a possible explanation for the convoluted relationship between the two diseases [2,3]. To uncover the promoting and protective factors of microcephaly with confirmed ZIKV (m-ZIKV+), we perform a large-scale network-based analysis of 382 non-redundant factors over each of the 5,665 municipalities in Brazil.

2 Method

In our network, nodes represent the 382 non-redundant factors and edges connect factors that are statistically significantly correlated over Brazil's municipalities (with partial correlation p-value < 0.05). As this methodology provides us with a very dense network that is the closes to an Erdos-Renyi random graph, we further threshold the network until it assumes a non-random topology (see Figure 1).

In our approach, the influence patterns around a particular node in the network are captured by the subnetwork induced by the considered node and by its direct neighbors. In this way, we generate four context subnetworks: the subnetwork centered at m-

adfa, p. 1, 2011.

ZIKV+, which we compare to the subnetwork centered at microcephaly without m-ZIKV (m-ZIKV-), as well as with the sub-networks centered at Low Birth Weight (LBW) and all Births (B), which serve as a proxy for general birth defects and for healthy births, respectively.

Then, we compare m-ZIKV+ context subnetwork to the three other subnetworks according to: 1) the overlap of their node-sets to determine if different contexts are influenced by the same factors (nodes) ; 2) according to their edge-weights to determine if factors have (dis)similar protective/promoting effects on different contexts,;and 3) according to the subnetwork topological similarity to fully account for the interplay between factors influencing the different contexts. We compare the node sets of each pair of subnetworks applying the McNemar statistic. The edge-weights are compared using Spearman's rank correlation. The topological similarities are uncovered using GCD58 distance measure [4]. For GCD58 comparisons, an empirical p-value is estimated by the distribution of GCD58 distances between m-ZIKV$^+$ subnetwork and 200 subnetworks that are randomly generated from the entire network thresholded as described above by selecting a random node and its direct neighbors.

3 Results and discussion

All three methods of comparison indicate that the incidence patterns of m-ZIKV$^+$ and m-ZIKV$^-$ are related. The McNemar statistic's p-value of 0.51 indicates that there is no statistically significant difference in the set of variables influencing m-ZIKV$^+$ and m-ZIKV$^-$. Furthermore, a correlation of 99% between the edge values adds that these variables affect m-ZIKV$^+$ and m-ZIKV$^-$ in a statistically significantly similar way, both in terms of sign as in in terms of value. These findings are further strengthened as, according to the GCD58 network distance measure, the structure of the incidence patterns of m-ZIKV$^+$ and m-ZIKV$^-$ are statistically significantly similar at the empirical p-value of 5%. In conclusion, these findings show that the incidence pattern related to m-ZIKV$^+$ is not specific to ZIKV infection and is therefore common to microcephaly with or without ZIKV. It is important to note however, that this finding could be explained by a diagnostic failure to detect ZIKV infection, in turn implying a possible underestimation of the ZIKV outbreak. Note that our network model allows for the precise identification of m-ZIKV$^+$ influencing variables, which is subject of further research.

The incidence pattern of m-ZIKV$^+$ is however statistically significantly different from that of LBW in terms of variables and topology. Nevertheless, a statistically significant correlation between the subset of shared associations (i.e. edge-weights) indicates that the variables common to both incidence patterns influence m-ZIKV$^+$ and LBW in the same way, both in terms of relative size and in terms of sign.

As expected, the incidence pattern of m-ZIKV$^+$ is statistically significantly dissimilar to that of "All Births" in terms of nodes and topology. Furthermore, the edge-weights are actually statistically significantly reversely correlated with the p-value of 2.09 e -05. This is to be expected, as it means that variables stimulating m-ZIKV$^+$ have a negative effect on the number of healthy births and vice versa.

4 References

1. COES, "Informe Epidemiológico N° 32 - SE 25/2016. Monitoramento dos casos de microcefaliano Brasil" (2016), (available at http://combateaedes.saude.gov.br/images/pdf/informe_microcefalia_epidemiologico_32.pdf).
2. D. Butler, Brazil asks whether Zika acts alone to cause birth defects. *Nature*. **535** (2016), pp. 475–476Ö.
3. Evans, D., Nijhout, F., Parens, R., Morales, A. J., & Bar-Yam, Y. (2016). A possible link between pyriproxyfen and microcephaly. arXiv preprint arXiv:1604.03834.
4. N. Yaveroğlu et al. (2014), Revealing the hidden language of complex networks. Scientific reports 4, 4547.
5. Newman, M. E. (2002). Random graphs as models of networks. arXiv preprint cond-mat/0202208.
6. Penrose, M. (2003). Random geometric graphs (No. 5). Oxford University Press.
7. Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509-512.
8. Pržulj, N., & Higham, D. J. (2006). Modelling protein–protein interaction networks via a stickiness index. Journal of the Royal Society Interface, 3(10), 711-716.

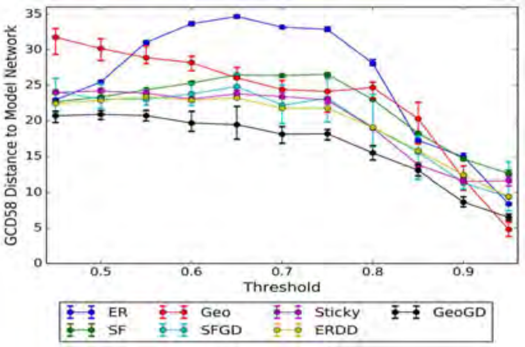


Fig. 1. Further thresholding of the network to achieve a non-random topology. We threshold the edges at different minimum levels of absolute (partial) correlation and subsequently fit the resulting thresholded network to model networks using *GCD58* [2]. Model networks fitted are: Erdős–Rényi (ER) [5], Scale Free (SF) [*7*], Geometric (Geo) [*6*], Scale-Free Gene Duplication (SFGD) *[7*], Sticky [9], Erdős–Rényi with Degree Distribution of the data [5] and Geometric with Gene Duplication (GeoGD) [*6*]).

Complex network analysis of images of human retina

Pablo Amil[1], Irene Sendiña[2], and Cristina Masoller[1]

[1] Universitat Politecnica de Catalunya, Barcelona, Spain
pa1000kmph@gmail.com,
[2] Universidad Rey Juan Carlos, Madrid, Spain

1 Introduction

Fundus images are color images taken from the posterior part of the eye (the retina) by means of coupling an optical apparatus with the eye optics. In this work we apply complex network tools to fundus images with the aim of automatically recover all the network structural information, which can yield information that can be useful to diagnose eye diseases.

To recover the structural information we first perform several filtering process to the original images in order to enhance the contrast between the vessel network and the retina. Then, we use a graph-based segmentation algorithm [1], and finally we perform several morphological operations to the segmented image. The result of such operations is the adjacency matrix that is then analyzed to recover all the network structural information.

2 Results

We analyzed 45 fundus images from a publicly available database [2] which are divided in 3 groups: healthy, glaucoma, and diabetes. These images have a resolution of 3504-by-2336 pixels. An example is shown in figure 1, left. In order to enhance the contrast between the vessel network and the retina, we first perform a filtering processes. Then, by running a segmentation algorithm (adapted from Ref.[1]) a new image is obtained from where a list of nodes (bifurcation points and endpoints) with their locations can be extracted, as well as the path connecting the nodes. An example of the resulting image is shown in figure 1, right.

From this information an adjacency matrix is obtained which is then used for characterization and analysis. Three sets of features can be extracted and are being analyzed. The first set is obtained from each individual network: simple features include the number of nodes and the number of links; advanced ones are based in information-theory and characterize the degree distribution and the distance distribution (the Shannon entropy, the statistical complexity, and the Fischer information [3,4]). A second set of features can be extracted from the comparison of pairs of networks. For example, a nonlinear dimensionality reduction algorithm (*IsoMap* [5]) is applied to the set of Jensen-Shannon distances between the degree distributions, or to the set of dissimilarity distances computed as in Ref.[6]. A third set of features can be extracted by comparing real networks with synthetic networks; in this case the features extracted are the parameters of the algorithm (e.g. [7]) that generates the tree-like structure that is more similar to the real network. Ongoing work is devoted to analyze this large set of features in order to determine which ones allow to classify the images in groups that more closely reflect the underlying ophthalmology classification.

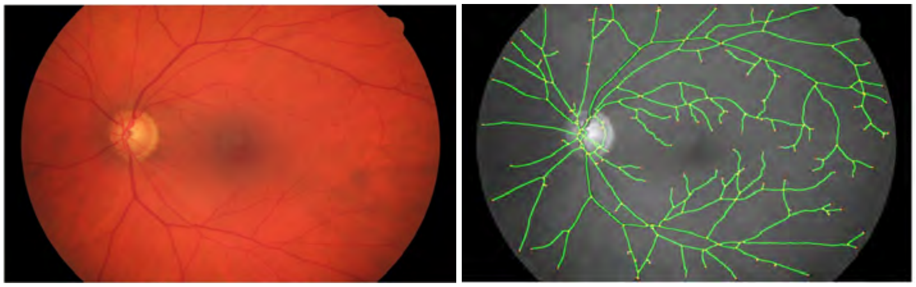


Fig. 1. Left: original fundus image; right: after filtering and segmentation, a tree-like network is extracted.

Summary. We use network tools to characterize retina fundus images. The resulting network can have information that can be useful to diagnose eye diseases.

Acknowledgment. This work is supported by the European Union through the project Advanced Biomedical Optical Imaging and Data Analysis (ITN BE-OPTICAL) under Grant 675512.

References

1. D. Santos Sierra, I. Sendiña-Nadal, I. Leyva, J.A. Almendral, A. Ayali, S.Anava, C.Sanchez-Avila, and S. Boccaletti: Graph-based unsupervised segmentation algorithm for cultured neuronal networks' structure characterization and modeling. Cytometry 87A, 513–523 (2015).
2. Erlangen Nurenberg Friedrich-Alexander Universitat. High-resolution fundus (hrf) image database, 2001.
3. O. A. Rosso, R. Ospina, and A. C. Frery: Classification and verification of handwritten signatures with time causal information theory quantifiers. PLOS ONE 11, e0166868 (2016).
4. M. Wiedermann, J. F. Donges, J. Kurths, and R. V. Donner: Mapping and discrimination of networks in the complexity-entropy plane. Arxiv:1704.07599v1 (2017).
5. J. B. Tenenbaum, V. De Silva, and J. C. Langford: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000).
6. T. A. Schieber, L. Carpi, A. Diaz-Guilera, P. M Pardalos, C. Masoller, and M. G. Ravetti: Quantification of network structural dissimilarities. Nat. Comm. 8, 13928 (2017).
7. N. Vandewalle and M. Ausloos: Construction and properties of fractal trees with tunable dimension: The interplay of geometry and physics. Phys. Rev. E 55, 94–98 (1997).

Investigation of control profile in biological networks

Vandana Ravindran[1], V Sunitha[1], and Ganesh Bagler[2]

[1] Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India.
[2] Center for Computational Biology, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), New Delhi, India.

1 Introduction

Studying control properties of a complex network provides insights into how the dynamical system represented by the network can be influenced to achieve desired behaviour. Many biological systems which are important for the normal regulation of a cell are dynamic systems. Undesirable behaviour of such systems is observed in the form of diseases and has lead to interest in studying the control of such complex systems through the corresponding networks. In our earlier work, we constructed and analysed the human cancer signalling network and established the connection between driver nodes which when provided with an external input can trigger the system to a desired state and their implications on cancer [2]. The maximum matching model implemented therein have shown the minimum number of controls required to control the system [1]. While topological properties of a network like its degree distribution is correlated with the minimum number of controls, it does not provide an explanatory detail of each control. For example a financial system and a biological system may require the same number of controls, but the structures giving rise to these controls may be different. Understanding the control properties of a complex network requires more than just knowing the number of controls. For an effective control strategy, it is also important to characterize the functional origin of each control. To know why a node is a driver node, we look at the control profile of the network. It is a statistic that quantifies the different proportions of control-inducing structures present in the network [3].

In this work, we determine why a node is a driver node based on its position in the network, and identify the control profile for some biological networks and deduce the control strategy for each of these networks.

2 Results

We analysed few biological networks and identified the control profile for each of them based on the method given by Ruth *et.al.* [3]. To understand why a node is a driver node, we decompose the driver nodes into three groups [3]: (1) *source nodes*- appear

vandana_ravindran@daiict.ac.in,
v_suni@daiict.ac.in
bagler@iiitd.ac.in

at the origin of the stem and they must be directly controlled, (2) *external dilations* which arise due to surplus of sink nodes (nodes that have no out-going links)- since each source node can control only one sink node, the number of external dilation N_e is $max(0, N_t - N_s)$, and (3) *internal dilations*- a structure that occurs when a path branches into two or more paths in order to reach other nodes; N_i denotes the number of internal dilations in a network. Thus the minimum number of independent controls N_D required to gain full control is the sum of the number source nodes, the external dilations and the internal dilations i.e. $N_D = N_s + N_e + N_i$. The control profile for a network is given by $(\eta_s = N_s/N_D, \eta_e = N_e/N_D, \eta_i = N_i/N_D)$, where, N_D is the number of driver nodes and is computed using the maximum matching model for directed graphs [1]. It is the set of those nodes that are never matched in the matching algorithm.

The table below summarizes the control profile for the networks analysed (Table 1).

Network	Nodes	Edges	Driver nodes	η_s	η_e	η_i
Cancer Signalling	1232	3060	47%	0.29	0	0.18
Directed Human PPI	6339	34813	36%	0.06	0.02	0.28
HIV-human molecular	6361	40625	36%	0.04	0.04	0.28
T-cell activation	121	255	29%	0.13	0	0.16
HIV- T-cell activation	137	367	25%	0.10	0	0.15
E.coli transcription	423	578	73%	0.11	0.62	0.002

Table 1. Control profile of some biological networks

We analyse the control profile of some biological networks and characterise them based on whether the network is dominated by source nodes, or by external dilations or by internal dilations. The human cancer signalling network, is a signal transduction network that is characterised with cancer associated genes and pathways altering cancer [2]. This network is source dominated (Table 1). This means the ratio of sinks to sources is less than one i.e., there are fewer sinks than sources. These networks have no external dilation. But this does not mean that there are no sink nodes. It means there is at least one distinct source which reaches a sink through a directed path. Since the source nodes lie at the boundary they are easily accessible and therefore are control targets. For example, receptor proteins responsible for transducing extracellular stimuli into intracellular signals which were characterised as PDNs (Peripheral driver nodes) in the human cancer signalling network are the source nodes [2]. Since these source nodes are readily accessible and influence the protein interactions within the cell, they can be used as potential drug targets [4]. This procedure which is in practise, is in synchronisation with the results of our theoretical study.

The T-cell and HIV-1 T-cell activation networks are networks that describe the T-cell activation pathway without and with HIV-1 infection. These networks are internal dilation dominated. They have no external dilation (Table 1). In such signalling networks, the source and certain intra-cellular molecules drive the signal transduction within the cell. For instance the HIV-1 virus attacks the CD4 receptor which in turn activates other down-stream proteins, for the release of T-helper cells. Thus the source and some internal dilation nodes are responsible for control. The Human protein-protein (PPI) network is a directed protein-protein interaction network. The HIV-1 human molecular network represents the HIV-1 interactions with human proteins upon infection. These networks are also internal dilation dominated networks (Table 1). Such networks lack sources which indicate clear input and sinks that indicate clear system output. This means that the system is closed or mostly closed [3]. The PPI network and the HIV-1 molecular network also have feedback loops and forms a closed system.

The *E.coli* transcription network, which is a gene regulatory network is external-dilation dominated (Table 1). This implies that sinks outnumber the source nodes. As a result controls applied to sources will yield correlated behaviour within the network. The genes in a transcription network exhibit high degree of correlation expression. For instance a particular gene can co-express or down-regulate the expression of another gene. Thus, if we seek to fully control such a system, we need to add controls beyond the sources.

3 Conclusion

The dynamical properties of a cell are hardwired in the genome and influenced by environmental and epigenetic changes. Thus the cell is naturally receptive to external cues and this provides us an opportunity for its manipulation to achieve desired outcomes. In order to take the best advantage of this property we require a deeper understanding of when and where to apply these external influences. By looking at the control profile of networks, one can get a better understanding of the ease with which the network can be controlled and the nature of driver nodes. Control profiles offer a way to capture the origin of control in networks and to understand why a node acts as a driver node. Studying control profile of biological networks helps in improved understanding of the system and in identifying those nodes that can efficiently control the system. For the networks analysed, we can deduce, based on the control profile that the source dominated networks like human cancer signalling and T-cell signalling network can be efficiently controlled through the source nodes. In the case of internal dilated networks like HIV-human molecular network and the human PPI network which are closed systems and obey certain conservation laws, some non-source nodes are also required to gain control. We conclude that the control profile of a network adds insights about the structure of the network which could then offer strategic ways to control the system. This is helpful particularly in biological networks that are highly constrained and require large sets of control nodes to gain full control.

References

1. Liu, Y.Y., Slotine, J.J., Barabási, A.L.: Controllability of complex networks. Nature 473(7346), 167–173 (2011)
2. Ravindran, V., Sunitha, V., Bagler, G.: Identification of critical regulatory genes in cancer signaling network using controllability analysis. Physica A: Statistical Mechanics and its Applications 474, 134–143 (2017)
3. Ruths, J., Ruths, D.: Control profiles of complex networks. Science 343(6177), 1373–1376 (2014)
4. Williams, M., Raddatz, R.: Receptors as Drug Targets. John Wiley and Sons, Inc. (2001), http://dx.doi.org/10.1002/0471141755.ph0101s32

Part III

Brain Networks

Coalescent embedding in the hyperbolic space unsupervisedly discloses the hidden geometry of the brain

Alberto Cacciola[1], Alessandro Muscoloni[2], Vaibhav Narula[2], Alessandro Calamuneri[3], Salvatore Nigro[4], Emeran A. Mayer[6,7,8,9], Jennifer S. Labus[6,7,8], Giuseppe Anastasi[3], Aldo Quattrone[4,5], Angelo Quartarone[1,3], Demetrio Milardi[1,3] and Carlo Vittorio Cannistraci[1,2,*]

[1] IRCCS Centro Neurolesi “Bonino Pulejo”, Messina, Italy.
[2] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[3] Department of Biomedical, Dental Sciences and Morphological and Functional Images, University of Messina, Messina, Italy.
[4] Institute of Bioimaging and Molecular Physiology, National Research Council, Catanzaro, 88100, Italy.
[5] Institute of Neurology, Department of Medical and Surgical Sciences, University "Magna Graecia", Catanzaro, 88100, Italy.
[6] G. Oppenheimer Center for Neurobiology of Stress and Resilience, UCLA, Los Angeles, CA, United States
[7] Department of Medicine, UCLA, Los Angeles, CA, United States
[8] UCLA Vatche and Tamar Manoukian Division of Digestive Diseases, UCLA, Los Angeles, CA, United States
[9] UCLA Brain Research Institute, Los Angeles, CA, United States

The human brain displays a complex network topology, whose structural organization is widely studied using diffusion tensor imaging [1], [2]. The original geometry from which emerges the network topology is known, as well as the localization of the network nodes in respect to the brain morphology and anatomy. One of the most challenging problems of current network science is to infer the latent geometry from the mere topology of a complex network. The human brain structural connectome represents the perfect benchmark to test algorithms aimed to solve this problem. Coalescent embedding [3], [4] was recently designed to map a complex network in the hyperbolic space, inferring the node angular coordinates. Here we show that this methodology is able to unsupervisedly reconstruct the latent geometry of the brain with an incredible accuracy and that the intrinsic geometry of the brain networks strongly relates to the lobes organization known in neuroanatomy. Furthermore, coalescent embedding allowed the detection of geometrical pathological changes in the connectomes of Parkinson’s Disease patients. The present study represents the first evidence of brain networks’ angular coalescence in the hyperbolic space, opening a completely new perspective, possibly towards the realization of latent geometry network markers for evaluation of brain disorders and pathologies.

References

[1] E. Bullmore and O. Sporns, “Complex brain networks: graph theoretical analysis of structural and functional systems,” *Nat. Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, 2009.

[2] D. S. Bassett and O. Sporns, “Network neuroscience,” *Nat. Neurosci.*, vol. 20, no. 3, p. 353, 2017.

[3] J. M. Thomas, A. Muscoloni, S. Ciucci, G. Bianconi, and C. V. Cannistraci, “Machine learning meets network science: dimensionality reduction for fast and efficient embedding of networks in the hyperbolic space,” *arXiv:1602.06522*, 2016.

[4] A. Muscoloni, J. M. Thomas, S. Ciucci, G. Bianconi, and C. V. Cannistraci, “Machine learning meets complex networks via coalescent embedding of networks in the hyperbolic space,” *Nat. Commun.*, 2017 *(accepted for publication)*.

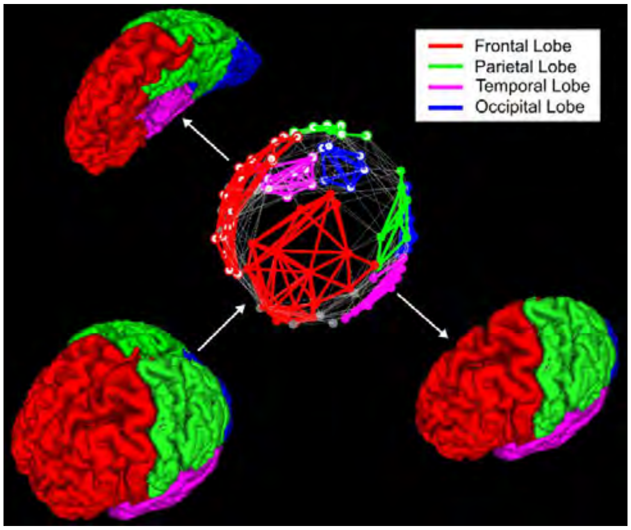


Fig. 1. The average (median) structural connectivity matrix of 30 healthy controls (HC) has been mapped in the 3D hyperbolic space using the coalescent embedding ISO technique. The figure shows, in a superior-anterior-lateral view, the 3D geometry of the brain emerging from the embedding in the hyperbolic sphere. The colours-filled circles represent the nodes of the left hemisphere, whereas the white-filled ones represent the brain structures of the right hemisphere. Each node has been labelled according to its real anatomical localization in the different brain lobes (see legend), the grey colour characterizes the nodes that have not been assigned to any lobe, since they represent grey matter structures placed in the deep white matter. It is worthy to note that the brain network geometry resembles almost perfectly the real brain anatomy, as evident from the 3D representation of a real brain. The whole brain placed anteriorly to the reconstructed network has been split into the left and right hemispheres in order to show the Right Temporal (magenta) and Occipital (blue) Lobes and to make even more visible the close relation between the latent geometry of the brain and the brain anatomy itself. Furthermore, another interesting finding is that we were able to reconstruct such latent geometry unsupervisedly starting from the mere topology of the network.

	mean marker (HC)	**mean marker (PD)**	**MW p-value**	**AUC**	**AUPR**
MCE-HSP	14.7	15.7	**0.006**	0.87	0.82
ncMCE-HSP	14.3	15.2	**0.014**	0.83	0.79
MCE-HD	14.4	15.3	**0.017**	0.82	0.77
ncMCE-HD	14.1	14.9	**0.017**	0.82	0.77
LE-HSP	15.6	16.3	**0.017**	0.82	0.78
ncISO-EA-HSP	15.9	16.7	**0.021**	0.81	0.78
ncMCE-EA-HSP	15.9	16.7	**0.021**	0.81	0.78
ISO-EA-HSP	15.9	16.7	**0.026**	0.80	0.77
ncISO-HSP	15.5	16.2	**0.026**	0.80	0.76
MCE-EA-HSP	15.9	16.7	**0.026**	0.80	0.77
LE-EA-HSP	15.9	16.7	**0.026**	0.80	0.77
LE-HD	15.0	15.6	**0.038**	0.78	0.72
ISO-HD	14.8	15.4	**0.045**	0.77	0.71
ISO-HSP	15.4	16.2	**0.045**	0.77	0.73
ncISO-HD	14.9	15.5	0.121	0.71	0.64
ncISO-EA-HD	15.2	15.7	0.121	0.71	0.69
LE-EA-HD	15.2	15.7	0.121	0.71	0.69
ISO-EA-HD	15.2	15.7	0.186	0.68	0.66
MCE-EA-HD	15.2	15.7	0.186	0.68	0.66
ncMCE-EA-HD	15.2	15.7	0.186	0.68	0.66
weights	222.0	215.8	0.307	0.64	0.64

Table 1. The connectomes of 10 de novo drug naïve Parkinson's Disease (PD) patients and 10 age- and sex-matched Healthy Controls (HC) have been mapped using the coalescent embedding algorithms. The networks represent the Number of Streamlines (NOS) between pairwise regions as provided by Constrained Spherical Deconvolution (CSD) tractography. To each subject has been assigned a geometrical marker equal to either the average hyperbolic distance (HD) or hyperbolic shortest path (HSP) between the embedded nodes. The table reports for each method the mean marker of the two groups. For reference, the average edge weight in the original network has been also used as a marker. The discrimination between the two groups has been assessed according to different metrics: the p-value of the Mann-Whitney (MW) test, the area under the receiver operating characteristic curve (AUC) and the area under precision-recall curve (AUPR). Note that significant p-values are highlighted in bold, considering a confidence level of 0.05. As emerging from the table, almost all the methods uncover a significantly different geometry underlying the human structural brain networks in de novo drug naïve PD patients.

Filtering information in complex brain networks

Fabrizio De Vico Fallani[1,2], Vito Latora[3,4], and Mario Chavez[2]

[1] Inria Paris, Aramis project-team, Paris, France
fabrizio.devicofallani@gmail.com
[2] CNRS UMR-7225, Sorbonne Universites, UPMC Univ Paris 06, Inserm, Institut du cerveau et de la moelle (ICM) - Hopital Pitie-Salpetriere, Paris, France
[3] School of Mathematical Sciences, Queen Mary University of London, London, UK
[4] Dipartimento di Fisica e Astronomia, Universita di Catania and INFN, Catania, Italy

1 Extended abstract

Complex brain networks are mainly estimated from empirical measurements. As a result of the inference process, we obtain a matrix of values corresponding to a fully connected and weighted network. To turn this into a useful sparse network, filtering procedures are typically adopted to prune weakest connections [?]. However, network properties strongly depend on the number of remaining links and how to objectively fix a connectivity threshold is still an open issue [?].

Here, we propose a criterion (ECO) to filter connectivity based on the optimization of fundamental properties of complex systems, ie efficiency and economy [?,?]. We prove analytically, and we confirm through numerical simulations, that the connection density maximizing such trade-off scales with the number of nodes according to a power-law (Fig. 1). This optimal density gives sparse networks (average node degree $k \sim 3$) yet emphasizing the intrinsic structural properties. We validate this result on several brain networks and we show the potential in discriminating neural diseases.

We suggest that ECO can advance our ability to analyze and compare biological networks, inferred from experimental data, in a fast and principled way.

References

1. Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. Neuroimage. 52(3):1059-69 (2010)
2. De Vico Fallani F, Richiardi J, Chavez M, Achard S. Graph analysis of functional brain networks: practical issues in translational neuroscience. Philos Trans R Soc Lond B Biol Sci. 369:1653 (2014)
3. Latora V, Marchiori M. Economic small-world behavior in weighted networks. Eur. Phys. J. B 32, 249-263 (2003)
4. Bullmore E, Sporns O. The economy of brain network organization. Nat Rev Neurosci. 13(5):336-49 (2012)

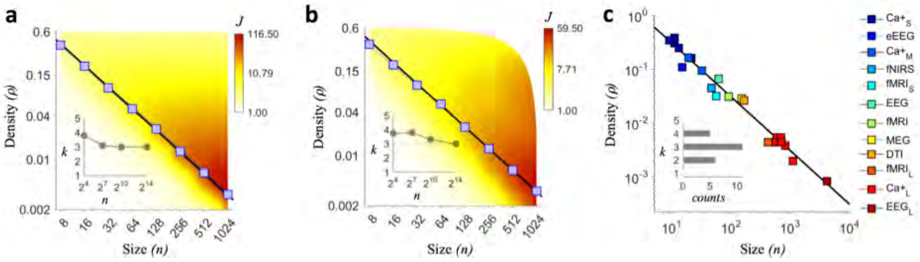


Fig. 1. Density threshold in synthetic networks and in brain networks. (a-b) Blue squares spot out the average connection density (ρ) values returned by the maximization of the efficiency-cost trade-off given by the equation $J = (Eglob + Eloc)/\rho$. The black line shows the fit $\rho = c/(n-1)$ to the data, with $c = 3.258$ for small-world networks and $c = 3.215$ for scale-free networks. The background color codes for the average value of the quality function J. Insets indicate that the optimal average node degree (k), corresponding to the density that maximizes J, converges to $k = 3$ for large network sizes ($n = 16834$). (c) Optimal density values maximizing group-averaged J profiles for different brain networks obtained from disparate neuroimaging data. The fit $\rho = c/(n-1)$ to the pooled data gives $c = 3.06$ (adjusted $R^2 = 0.994$). The inset shows a sharp distribution for the corresponding average node degree, with a mode $k = 3$.

Multiscale mixing patterns in networks

Leto Peel[1 2], Jean-Charles Delvenne[1 3] and Renaud Lambiotte[2 4]

[1] ICTEAM, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[2] naXys, Université de Namur, Namur, Belgium
[3] CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[4] Mathematical Institute, University of Oxford, Oxford, UK

Assortative mixing in networks is the tendency for nodes with the same attributes, or metadata, to link to each other. For instance in social networks we may observe more interactions between people with the same age, race, or political belief. Quantifying the level of assortativity or disassortativity (the preference of linking to nodes with different attributes) can shed light on the factors involved in the formation of links in complex networks. It is common practice to measure the level of assortativity according to the assortativity coefficient [2], which for discrete-valued attributes compares the proportion of links connecting nodes of with same attribute value, or *type*, relative to the proportion expected if the edges in the network were randomly rewired. The difference between these proportions is commonly known as modularity Q, a measure frequently used in the task of community detection [3]. The assortativity coefficient is a normalised such that $r_{\text{global}} = 1$ if all edges only connect nodes of the same type (i.e., maximum modularity $Q_{\max}$) and $r_{\text{global}} = 0$ if the number of edges is equal to the expected number for a randomly rewired network in which the total number of edges incident on each type of node is held constant. Mathematically, we calculate the global assortativity r_{global} by [2]

$$r_{\text{global}} = \frac{Q}{Q_{\max}} = \frac{\sum_g e_{gg} - \sum_g a_g^2}{1 - \sum_g a_g^2} \quad , \tag{1}$$

where e_{gh} is the proportion of edges in the network that connect vertices with type $y_i = g$ to vertices with type $y_j = h$ and a_g and b_g represent the total number of outgoing and incoming links of all nodes of type g:

$$e_{gh} = \frac{1}{2m} \sum_{ij} A_{ij} \delta_{y_i,g} \delta_{y_j,h} \quad , \qquad a_g = \sum_h e_{gh} \quad . \tag{2}$$

The global assortativity is a summary statistic that describes the pattern of mixing on average across the whole network. But as with all summary statistics there may be cases where it provides a poor representation of the network as a whole, for instance, if there are localised heterogeneous patterns of mixing across the network. Figure 1 illustrates an analogy to Anscombe's quartet of bivariate datasets with identical correlation coefficients [1]. Each of the five networks in the top row have the same number of nodes ($n = 40$) and edges ($m = 160$) and have been constructed to have the same r_{global} with respect to a binary attribute, indicated by a cross (c) or a diamond (d). All five networks have $m_{cc} + m_{dd} = 80$ edges between nodes of the same type and $m_{cd} = 80$ edges between nodes of different types, such that each has $r_{\text{global}} = 0$. Local patterns of mixing are formed by splitting each of the types $\{c, d\}$ further into two equally sized subgroups

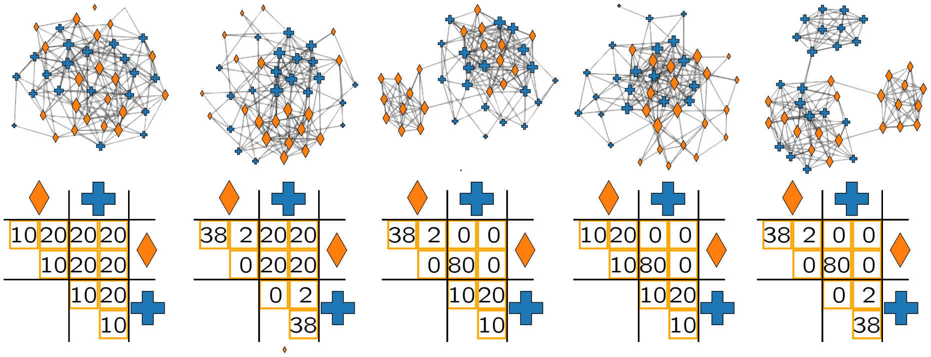


Fig. 1. Five networks (top) of $n = 40$ nodes and $m = 160$ edges with the same global assortativity $r_{\text{global}} = 0$, but with different local mixing patterns.

$\{c_1, c_2, d_1, d_2\}$. The bottom row depicts the placement of edges within and between the four subgroups.

We propose a local measure of assortativity $r(\ell)$ that captures the mixing pattern within the local neighbourhood of a given node ℓ. We do so by reweighting the nodes in the edge count (Eq. (2)) based on the stationary distribution of a random walk with restart from ℓ. Figure 2 provides an example of our approach applied to a simple line network. Full details of the method are given in [4]. Consequently we are able to capture and evaluate the distribution of mixing patterns across the network. For example, Figure 3 shows the results of applying our method to the Facebook 100 dataset [5].

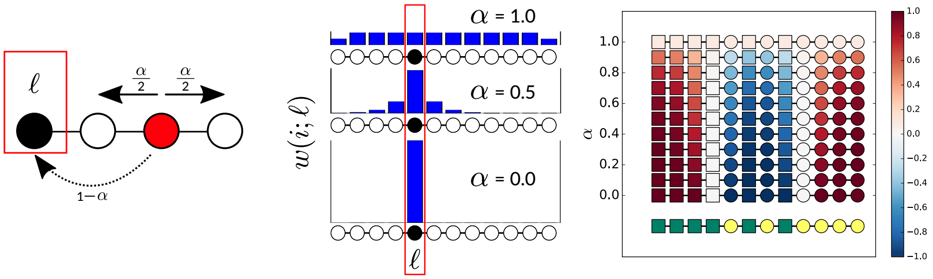


Fig. 2. Example of the local assortativity measure for discrete attributes (a) assortativity is calculated (as in (1)) according to the actual proportion of links in the network connecting nodes of the same type relative to the expected proportion of links between nodes of the same type, (b) the nodes in the network are weighted according to a random walk with restart probability of $1-\alpha$, (c) an example of the local assortativity applied to a simple line network with two types of nodes: yellow or green. The blue bars show the stable distribution ($w(i;\ell)$) of the random walk with restarts at ℓ for different values of α. Underneath each distribution the nodes in the line network are coloured according to their local assortativity value.

Through comparison with null models that preserve the global mixing pattern and degree distribution, we may assess the representativeness of the global assortativity. Our new approach provides a lens through which we can examine the variability of mixing patterns with respect to the graph structure and therefore identify whether outliers correspond to concentrated subgroups of connected nodes or more uniformly distributed individuals across the network. Using synthetic examples we describe cases of heterogeneous assortativity and demonstrate that for many real-world networks the global assortativity is not representative of the mixing patterns throughout the network.

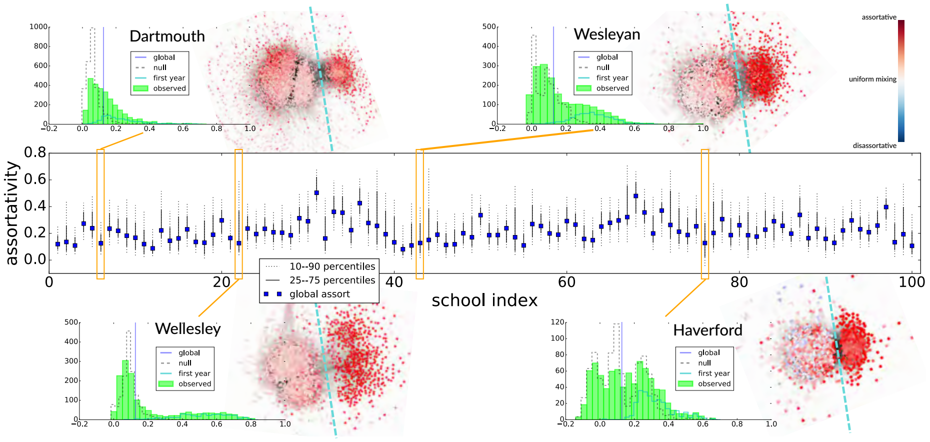


Fig. 3. Distributions of the local assortativity by residence (dorm) for each of the schools in the Facebook 100 dataset [5]. Dotted black lines indicate the 10 and 90 percentiles while the solid black lines show the interquartile range. The global assortativity is indicated by the blue square markers. The distributions for four schools (Dartmouth, Wesleyan, Wellesley and Haverford) are shown in detail in the surrounding. Each of them has approximately the same global assortativity ($r_{\text{global}} \sim 0.13$), but the distributions indicate different levels of heterogeneity in the pattern of mixing by residence. While the distributions are different, there exists a common trend that the first year students tend to be more loosely connected to the rest of the network and exhibit the higher values of assortativity (nodes to the right of the dashed cyan line).

Acknowledgements — Supported by IAP (Belgian Scientific Policy Office), ARC (Federation Wallonia-Brussels), FNRS and the MOVE-IN Louvain programme supported by the European Commission (COFUND Marie Curie actions).

References

1. Anscombe, F.J.: Graphs in statistical analysis. The American Statistician 27(1), 17–21 (1973)
2. Newman, M.E.: Mixing patterns in networks. Physical Review E 67(2), 026126 (2003)
3. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical review E 69(2), 026113 (2004)
4. Peel, L., Delvenne, J.C., Lambiotte, R.: Multiscale mixing patterns in networks. arXiv preprint arXiv:1708.01236 (2017)
5. Traud, A.L., Mucha, P.J., Porter, M.A.: Social structure of facebook networks. Physica A: Statistical Mechanics and its Applications 391(16), 4165–4180 (2012)

Regions of Interest as nodes of dynamic functional brain networks

Onerva Korhonen[1,2], Elisa Ryyppö[1], Enrico Glerean[3,2], Elvira Brattico[4], and Jari Saramäki[1]

[1] Department of Computer Science, Aalto University, Espoo, Finland,
onerva.korhonen@aalto.fi,
WWW home page: http://complex.cs.aalto.fi/
[2] Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland
[3] Turku PET Centre, University of Turku, Turku, Finland
[4] Center for Music in the Brain, Department of Clinical Medicine, Aarhus University, & The Royal Academy of Music Aarhus/Aalborg, Denmark

Information processing in the human brain is based on the balance between integration and segregation, *i.e.* clusters with strong internal connections and weaker inter-cluster connections [7, 6]. Therefore, it is natural to model the brain as a network, where nodes depict brain areas and links represent the structural or functional connections between these areas. Networks of the brain have hierarchical community structure, are organized into core and periphery, and have broad degree distributions and relative high global clustering and short path lengths between nodes [5, 1]. These networks are dynamic by nature: they change across the human lifespan and as well as on shorter timescales, for example between cognitive tasks [3, 6, 1].

The properties of functional brain networks strongly depend on how their nodes are defined. Commonly, so-called Regions of Interest (ROIs) are used as the nodes. ROIs are predefined collections of fMRI measurement voxels. The fMRI BOLD signal of each voxel reflects its activity, and the voxel signals inside a ROI are typically averaged to obtain a time series that represents the ROI. Therefore, in order to be a meaningful network node a ROI needs to be *functionally homogeneous*: the ROI time series is an accurate representation of the voxel dynamics inside the ROI only if each voxel has similar dynamics. Earlier, we have shown that ROIs of commonly used parcellations are not always functionally homogeneous [4]. For measuring homogeneity, we have used the *spatial consistency* that is defined as the average Pearson correlation coefficient between voxel time series inside a ROI.

Here, we ask how ROIs of the connectivity-based Brainnetome atlas [2] behave as nodes of dynamic brain networks. To this end, we divide the fMRI time series of 244 samples into five sliding windows of 80 samples. Window length may affect the obtained functional connectivity and spatial consistency since these are measured in terms of Pearson correlation coefficient. Here, we selected the window length so that the average spatial consistency calculated in the window does not significantly differ from the average spatial consistency calculated over the whole measurement time series.

To investigate ROIs as nodes of dynamic brain networks we use two measures: *spatiotemporal consistency* quantifies time-dependent changes in spatial consistency and *network turnover* measures the amount of turnover in the ROI's closest network neighborhood. We find that spatial consistency varies non-uniformly in space and time, which is reflected in the variation of spatiotemporal consistency across ROIs. Further, we see time-dependent changes in the ROIs' network neighborhoods, resulting in high network turnover. This turnover is non-uniformly distributed across ROIs, so that ROIs with high spatial and spatiotemporal consistency have low network turnover (Fig. 1).

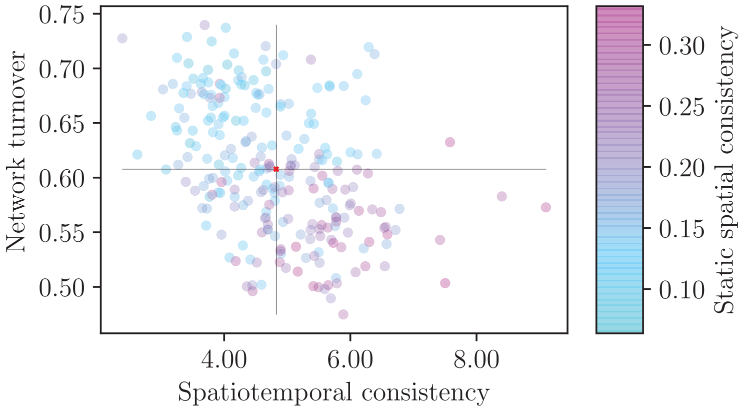


Fig. 1. Spatial and spatiotemporal consistency and network turnover are connected. ROIs with high spatial and spatiotemporal consistency tend to have low turnover in their network neighborhoods.

Finally, we reveal rich, time-dependent structure of voxel-level correlations inside ROIs (Fig. 2). One may speculate about the connections between this internal structure and the ROI's role in brain network topology and brain function. Would, for example, local and connector hubs differ in terms of their internal structure? The present tools of network neuroscience often ignore the internal connectivity of ROIs. For example, self-links are typically considered meaningless in brain networks; however, they could be used to represent changes in internal correlation structure of ROIs.

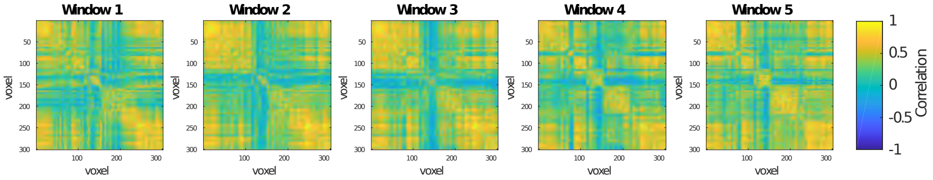


Fig. 2. Rich time-dependent structure of voxel-level correlations occurs inside ROIs. Pearson correlation matrices for left inferior frontal gyrus of a representative subject.

Because the internal structure and connectivity of ROIs vary in time, the common approach of using static node definitions may be surprisingly inaccurate. Instead, network neuroscience would greatly benefit from node definition strategies tailored for dynamic networks.

References

1. Basset, D.S., Sporns, O.: Network neuroscience. Nature Neuroscience 20(3), 353–364 (Feb 2017)
2. Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A.R., Fox, P.T., Eickhoff, S.B., Yu, C., Jiang, T.: The human Brainnetome atlas: a new brain atlas based on connectional architecture. Cerebral Cortex 26(8), 3508–3526 (Aug 2016)
3. Honey, C.J., Kötter, R., Breakspear, M., Sporns, O.: Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proceedings of The National Academy of Sciences (USA) 104(24), 10240–10245 (Jun 2007)
4. Korhonen, O., Saarimäki, H., Glerean, E., Sams, M., Saramäki, J.: Consistency of regions of interest as nodes of fMRI functional brain networks. Network Neuroscience (2017), advance online publication, doi: 10.1162/NETN_a_00013
5. Sporns, O.: The human connectome: origins and challenges. NeuroImage 80, 53–61 (Oct 2013)
6. Sporns, O.: Network attributes of segregation and integration in the human brain. Current Opinion in Neurobiology 23(2), 162–171 (Apr 2013)
7. Tononi, G., Sporns, O., Edelman, G.M.: A measure for brain complexity: relating functional segregation and integration in the nervous system. Proceedings of the National Academy of Sciences (USA) 91(11), 5033–5035 (May 1994)

Exponential random graph model for brain networks

Catalina Obando[1,2] and Fabrizio De Vico Fallani[1,2]

[1] Inria Paris, Aramis project-team, Paris, France
cobando85@gmail.com
[2] CNRS UMR-7225, Sorbonne Universites, UPMC Univ Paris 06, Inserm, Institut du cerveau et de la moelle (ICM) - Hopital Pitie-Salpetriere, Paris, France

1 Introduction

Network science has been extensively developed to characterize the structural properties of complex systems, including brain networks inferred from neuroimaging data. Quantifying the topological properties of brain networks by means of graph theory has reveled that they tend to exhibit similar organizational properties, including small-worldness, cost-efficiency, modularity and node centrality [1]. These results has enriched our understanding of the structure of functional brain connectivity maps, nevertheless, they refer to a descriptive analysis of the observed brain network, which is only one instance of several alternatives with similar structural features. Statistical models are, therefore, needed to reflect the uncertainty associated with a given observation, to permit inference about the relative occurrence of specific local structures and to relate local-level processes to global-level properties [5]. In this work we adopted a statistical model based on exponential random graphs (ERGM) to reproduce electroencephalographic (EEG) brain networks [6].

2 ERGM Model

Let G be a graph in a set $\mathscr{G}$ of possible network realizations, $g = [g_1, g_2, ..., g_r]$ be a vector of graph statistics, or metrics, and $g^* = [g_1^*, g_2^*, ..., g_r^*]$ the values of these metrics measured over G. Then, we can statistically model G by defining a probability distribution $P(G)$ over $\mathscr{G}$ such that the following conditions are satisfied:

$$\sum_{G \in \mathscr{G}} P(G) = 1 \tag{1}$$

$$\langle g_i \rangle = \sum_{G \in \mathscr{G}} g_i(G) P(G) = g_i^*, \quad i = \{1, 2, ..., r\} \tag{2}$$

where $\langle g_i \rangle$ is the expected value of the $i-th$ graph metric over $\mathscr{G}$.

By maximizing the Gibbs entropy of $P(G)$ constrained to the above conditions, the probability distribution reads as:

$$P(G) = \frac{e^{H(G)}}{Z} \tag{3}$$

where $H(G) = \sum_{i=1}^{r} \theta_i g_i(G)$ is the graph Hamiltonian, θ_i is the $i-th$ model parameter to be estimated and $Z = \sum_{G \in \mathscr{G}} e^{H(G)}$ is the so-called partition function. The estimated value of a parameter θ_i indicates the change in the (log-odds) likelihood of an edge for a unit change in graph metric g_i. If the estimated value of θ_i is large and positive, the associated graph metric g_i plays an important role in explaining the topology of G more than would expected by chance.

We considered graph metrics reflecting the basic properties of complex systems such as hub propensity and transitivity in the network. Specifically, we focused on k-stars to model highly connected nodes (hubs) and k-triangles to model transitivity, where k refers to the order of the structures as illustrated in figure 1(a).

3 Data set

We used high-density EEG signals freely available from the online PhysioNet BCI database [4, 7]. EEG data consisted of 1 min resting state with EO and 1 min resting state with EC recorded from 56 electrodes in 108 healthy subjects. We used the spectral coherence [2] to measure functional connectivity (FC) between EEG signals of sensors i and j at a specific frequency band f which results in a set of connectivity matrices. The nodes of brain networks represent brain regions and are set by the senors; edges of brain networks represent functional connectivity and are set by the value of spectral coherence between two nodes. We then thresholded the values in the connectivity matrices to retain the strongest links in each brain network. We validated the results over different threshold values.

4 Results and Discussion

Results showed that the model including k-triangles and k-stars statistically reproduced the main properties of the EEG brain networks measured by global- and local-efficiency. We fitted the model to brain network for each subject, frequency band and conditions and then simulated 100 networks for each configuration that we used to positively test the goodness-of-fit of the model and cross-validated the results. In the cross-validation phase, we show that the model captured other important brain network properties as measured by the clustering coefficient, the characteristic path length and the modularity.

Furthermore, the fitted ERGM parameter values provided complementary information showing that clustering connections are significantly more represented from EC to EO.

These findings support the current view of the brain functional segregation and integration in terms of modules and hubs, and provide a statistical approach to extract new information on the (re)organizational mechanisms in brains.

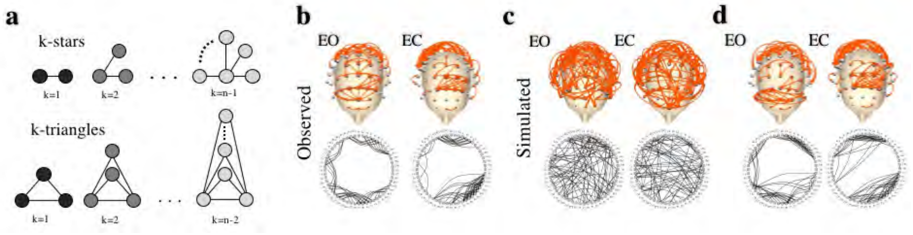


Fig. 1. (a) Graphical representation of *k*-stars and *k*-triangles. Among the different model configurations that we tested, the model including these two graph metrics was the one that ranked best in terms of relative errors between the mean values of global/local efficiency of the simulated networks and the value of the observed brain network. (b) Brain network in the *alpha* band for eyes-open (EO) and eyes-closed (EC) resting-state of one representative subject. (c) One instance of the corresponding synthetic networks sampled by the model. (d) Because node labels are not preserved in the simulated networks, we re-assigned them virtually by using the Frank-Wolfe algorithm [8], which optimizes the graph matching with the observed brain network. In the upper part of the figure, nodes correspond to EEG electrodes. In the bottom part, the nodes are arranged into a circle.

References

1. Bullmore, E. and Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nature reviews. Neuroscience, 10(3), p.186 (2009)
2. Carter, G.C.: Coherence and time delay estimation. Proceedings of the IEEE, 75(2), pp.236-255 (1987)
3. De VicoFallani, F., Richiardi, J., Chavez, M., Achard, S.: Graph analysis of functional brain networks: practical issues in translational neuroscience. Phil. Trans. R. Soc. B, 369(1653), 20130521 (2014)
4. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K. and Stanley, H.E.: Physiobank, physiotoolkit, and physionet. Circulation, 101(23), pp.e215-e220 (2000)
5. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U.: Network motifs: simple building blocks of complex networks. Science, 298(5594), pp.824-827 (2002)
6. Robins, G., Pattison, P., Kalish, Y., Lusher, D.: An introduction to exponential random graph (p*) models for social networks. Social networks, 29(2), 173-191 (2007)
7. Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N. and Wolpaw, J.R.: BCI2000: a general-purpose brain-computer interface (BCI) system. IEEE Transactions on biomedical engineering, 51(6), pp.1034-1043 (2004)
8. Vogelstein, J.T., Conroy, J.M., Lyzinski, V., Podrazik, L.J., Kratzer, S.G., Harley, E.T., Fishkind, D.E., Vogelstein, R.J. and Priebe, C.E.: Fast approximate quadratic programming for graph matching. PLOS one, 10(4), e0121002 (2015)

Mapping connectivity of bursting neuronal networks

Tuan Nguyen, Kelly O'Connor, Krishna Sheth, and Nick Bolle

Department of Physics, The College of New Jersey, Ewing NJ 08628, USA,
nguyena@tcnj.edu,
WWW home page: http://nguyen.pages.tcnj.edu/

1 Introduction

In dissociated culture, neuronal networks spontaneously generate synchronous activity known as network bursting: periodic firing of nearly all neurons every $30-100$ seconds. There has been much work towards understanding the initiation of bursts [1], the balance between excitatory and inhibitory connections [2], as well as the role of the underlying network topology [3]. Typically, assumptions have to be made regarding network connectivity or inferred from the correlated activity of individual neurons. Here we describe a new approach [4] that allows direct measurements of functional connectivity of a neuronal culture. The approach combines, for the first-time, laser scanning photostimulation (LSPS) of single neurons with simultaneous calcium (Ca) imaging of a large cell-population [5] to rapidly map excitatory connections in neuronal networks consisting of 150-200 neurons and 1500-2000 connections. Connectivity maps were measured throughout a 12-day period, in order to investigate network properties before and after the beginning of network bursting.

2 Apparatus and Methods

As shown in Fig. 1, single-neuron photostimulation utilizing caged-glutamate is achieved by focusing a 2-ms pulse of ultraviolet laser light to a spot $< 10\ \mu$m in diameter with a 4x microscope objective. Galvano-driven mirrors under computer control were utilized to rapidly steer the laser spot to any point within a 1350-μm$\times$1350-μm field-of-view (FOV) containing hundreds of neurons. For the Ca imaging component, 494-nm light from a high-power light emitting diode (LED) was directed through the same microscope objective. This light broadly illuminated the entire FOV in order to excite Fluo-4, a calcium indicator dye. Large-scale activity of all neurons in the FOV can be observed by detecting changes in the 516-nm emission fluorescence. Experiments were performed on primary cultures using cortical neurons dissociated from embryonic day 17 (E17) Sprague-Dawley rats. Final cell density was 300-400 cells/mm^2, which contained both excitatory and inhibitory neurons. At DIV 8-12, network bursting begins when $> 20\%$ of neurons fire together roughly every 2 minutes.

After recording ten minutes of spontaneous network activity, mapping began by first stimulating a randomly chosen neuron near the center of the FOV and identifying neurons that showed changes in fluorescence. Ca responses from neurons that were directly connected to the photostimulated neuron, i.e. first order neurons, began

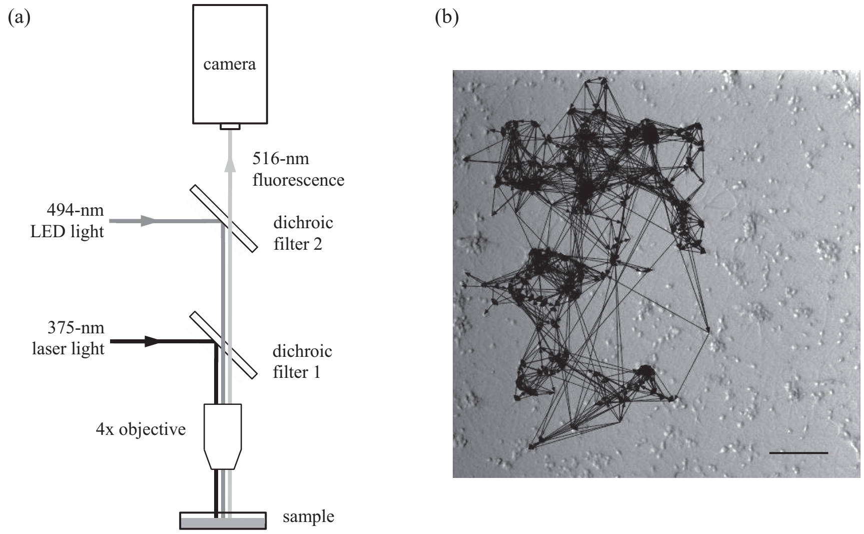


Fig. 1. (a) Schematic of apparatus used for simultaneous laser scanning photostimulation and calcium imaging.(b) Bright-field image of neuronal network (DIV 16) with large bursting activity combined with an overlay of corresponding measured connectivity map. Scale bar: 200 μm

at the same imaging frame as the response from the photostimulated neuron. While Ca responses from second order neurons could be elicited by increasing the light power and/or pulse duration, these had onset times of 3-5 s after photostimulation and could easily be distinguished. Because Ca responses at this level of sensitivity correspond to supratheshold events (i.e. action potentials), only excitatory connections were mapped. Within 2-3 hours, raw map data consisting of $\sim$ 200 neurons and $\sim$ 2000 connections were typically made. In one day, 1-3 networks of similar age, i.e. from same original batch, were usually mapped.

3 Results

From these map data, the adjacency matrix for an unweighted and directed network was constructed and used to calculate network properties. The average in-degree for each network increased from initial values of $\langle k_{in} \rangle_{avg} \simeq 4$ to a stable values of $\langle k_{in} \rangle_{avg} \simeq 10$ after the onset of network bursting [Fig. 2(a)]. Average clustering coefficients were calculated and compared to a randomized network model. We found that while clustering was much larger with $\langle C \rangle_{avg} \simeq 0.5$ than in the random network model [Fig. 2(b)], it was relatively constant over the entire period, and thus uncorrelated with network bursting. Figure 2(c) also shows the average global efficiency for each day rapidly increasing from an initial value of $\simeq 0.13$ to a steady value of 0.29 within four days (DIV 7-10) before the onset of network bursting. By contrast, global efficiency of the random network model stays relatively high throughout. The ‘small-world’ behavior of these networks was also assessed by considering the local efficiency, which is similar to the global efficiency but only applied to the node’s neighbors instead of the entire network.

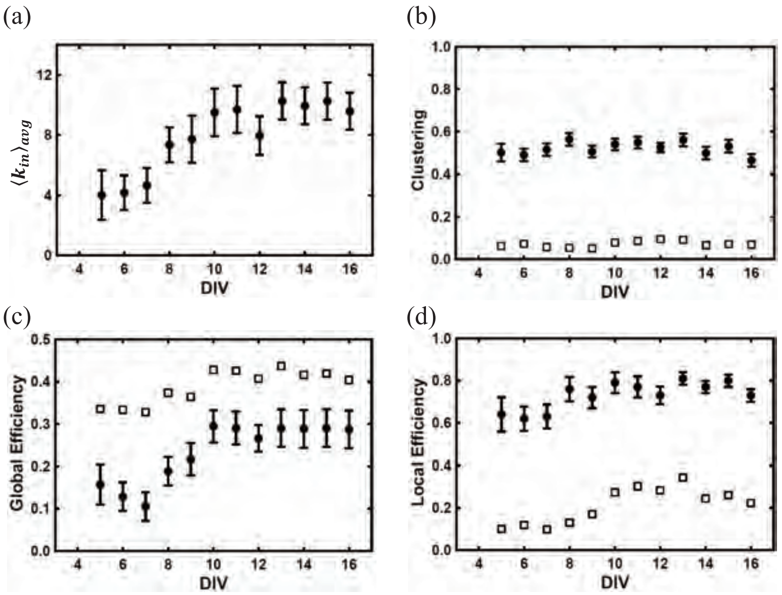


Fig. 2. Measured network properties. (a) Average in-degree, (b) clustering, (c) global efficiency, and (d) local efficiency as a function of neuronal age. Solid circles: averages of all networks mapped for given day. Open squares: averages computed from a randomized network model having the same degree distribution as each mapped network.

The initial local efficiency was large $\langle E_L \rangle_{avg} \simeq 0.63$ and increased by roughly 20% over time [Fig. 2(d)]. As discussed in [6], these networks exhibited small-world network behavior because both the global and local efficiency remained relatively large throughout the entire time period. As expected, the random network model did not show this behavior as its local efficiency remained relatively low.

References

1. Orlandi J., Soriano J., Alvarez-Lacalle E., Teller S., Casademunt J.: Noise focusing and the emergence of coherent activity in neuronal cultures. Nature Physics 9(9), 582–590 (2013)
2. Tibau E., Valencia M., Soriano J.: Identification of neuronal network properties from the spectral analysis of calcium imaging signals in neuronal cultures. Frontiers in Neural Circuits 7(199), 1–16 (2013)
3. Moriya S., Yamamoto H., Hisanao A., Hirano-Iwata A., Niwano, M., Kubota S., Sato S.: Modularity-dependent modulation of synchronized bursting activity in cultured neuronal network models. Proceedings of the International Joint Conference on Neural Networks (IJCNN) 1163–1168 (2017)
4. Nguyen T., O'Connor K., Sheth K., Bolle N.: Mapping functional connectivity of bursting neuronal networks. Applied Network Science 2(15), 1–15 (2017)
5. Cossart R., Ikegaya Y., Yuste R.: Calcium imaging of cortical networks dynamics. Cell Calcium 37(5), 451–7 (2005)
6. Latora V., Marchiori M.: Economic small-world behavior in weighted networks. Eur. Phys. Rev. J. B. 32(2), 249-263 (2003)

Loss of inter-frequency brain hubs in Alzheimer's disease

Jeremy Guillon[2,1], Yohan Attal[3], Olivier Colliot[2,1], Valentina La Corte[2], Bruno Dubois[4], Denis Schwartz[2], Mario Chavez[2,1], and Fabrizio De Vico Fallani[1,2]

[1] Inria Paris, Aramis project-team, 75013, Paris, France
[2] CNRS UMR-7225, Sorbonne Universites, UPMC Univ Paris 06, Inserm U-1127, Institut du cerveau et de la moelle épinière (ICM), Hôpital Pitié-Salpêtrière, 75013, Paris, France
[3] myBrain Technologies, Paris, France
[4] Department of Neurology, Institut de la Memoire et de la Maladie d'Alzheimer - IM2A, Paris, France

1 Introduction

Alzheimer's disease (AD) causes alterations of brain network structure and function. The latter consists of connectivity changes between oscillatory processes at different frequency channels. We proposed a multi-layer network approach to analyze multiple-frequency brain networks (Figure 1a) inferred from magnetoencephalographic recordings during resting-states in AD subjects and age-matched controls. We used the multi-participation coefficient (*MPC*) to quantify the tendency of brain regions to facilitate information propagation across different frequencies. Finally, we tested the diagnostic power of the measured brain network features to discriminate AD patients and healthy subjects.

2 Methods

The study involved 25 Alzheimer's diseased (AD) patients (13 women) and 25 healthy age-matched control (HC) subjects (18 women). All participants underwent the Free and Cued Selective Reminding Test (FCSRT) for verbal episodic memory. Specifically, we considered the Total Recall (TR) score - given by the sum of the free and cued recall scores - which has been demonstrated to be highly predictive of AD [4, 5].

We reconstructed the MEG activity on the cortical surface by using a source imaging technique [2]. The reconstructed time series were then averaged within 148 regions of interest (ROIs) defined by the Destrieux atlas. We estimated functional connectivity by calculating the spectral coherence between each pair of ROI signals. As a result, we obtained for each subject, a set of connectivity matrices of size 148×148 where the (i, j) entry contains the value of the spectral coherence between the signals of the ROI i and j at a frequency $f = 0, 0.5, \ldots, 499$. We then averaged the connectivity matrices within the following characteristic frequency bands [1]: *delta* ($2-4$ Hz), *theta* ($4.5-7.5$ Hz), *alpha1* ($8-10.5$ Hz), *alpha2* ($11-13$ Hz), *beta1* ($13.5-20$ Hz), *beta2* ($20.5-29.5$ Hz) and *gamma* ($30-45$ Hz). We finally thresholded and binarized the values of each connectivity matrix in order to obtain a connection density corresponding to an average node degree k of 12, belonging to a range of values typically used in brain networks analysis. We considered the local multi-participation coefficient [3] MPC_i to measure how evenly a node i is connected to the different layers of the multiplex. By construction, if nodes tend to concentrate their connectivity in one layer, the global multi-participation coefficient tends to 0; on the contrary, if nodes tend to have the same number of connections in every layer, the MPC value tends to 1 (Figure 1c). In the singular case where a node is disconnected in every layer, we assigned $MPC_i = 0$ to avoid indeterminate results.

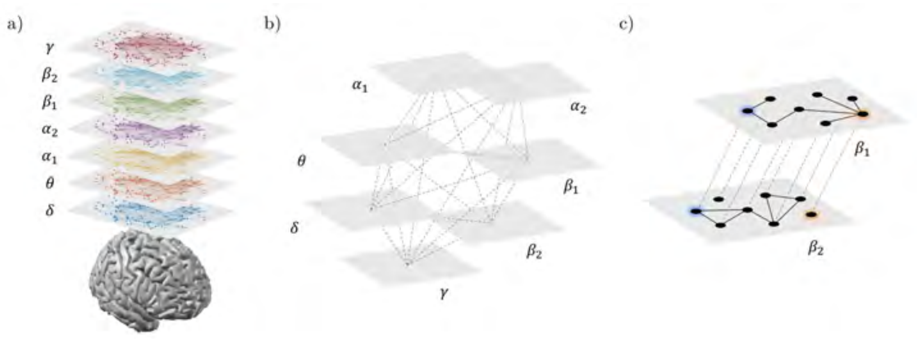


Fig. 1. Multi-frequency brain networks. Panel a) shows brain networks of a representative subject extracted from seven frequency bands. Links are inferred by means of spectral coherence and thresholded to have in each layer an average node degree $k = 12$. b) Procedure to construct a multi-frequency network. Each layer corresponds to a different frequency band. Only nodes representing the same brain region in each layer are virtually connected. Hence, inter-layer links code for identity relationships. c) Inter-frequency node centrality. A two-layer multiplex is considered for the sake of simplicity. The blue node acts as an inter-frequency hub (i.e., multi-participation coefficient $MPC = 1$) as it allows for a balanced information transfer between layer β_1 and β_2; the red node, who is disconnected in layer β_1, blocks the information flow and has $MPC = 0$.

3 Results

Main results showed that regional connectivity of AD subjects was abnormally distributed across the multiplex's layers (i.e. across frequency bands) as compared to controls (Figure 2b, Table 1).

MPC values significantly correlated with memory impairment of AD subjects, as measured by the free and cued selective reminding test (Figure 2a). Locally, the most predictive regions belonged to components of the default-mode network that are typically affected by atrophy, metabolism disruption and amyloid-β deposition.

We compared the diagnostic power between MPC and single-layer indices and we showed that the combination of the two types led to increased classification accuracy (78.39%) and sensitivity (91.11%).

Rank	ROI label	Cortex	Z score	p-value
1	G_precentral R	Motor	-3.4735	0.0006
2	G_front_inf-Opercular R	Motor	-2.5239	0.0127
3	S_oc_middle_and_Lunatus L	Occipital	-2.4582	0.0138
4	**G_pariet_inf-Supramar L**	Parietal	-2.4860	0.0142
5	**S_interm_prim-Jensen L**	Parietal	-2.3708	0.0147
6	**S_temporal_transverse R**	Temporal	-2.3996	0.0191
7	**S_pericallosal R**	Limbic	-2.3041	0.0203

Table 1. Statistical group differences for local multi-participation coefficient (MPC_i). ROI labels, abbreviated according to the Destrieux atlas, are ranked according to the resulting p-values. The same ranks are used as labels in Figure 2b. ROIs highlighted in bold belong to the default mode network.

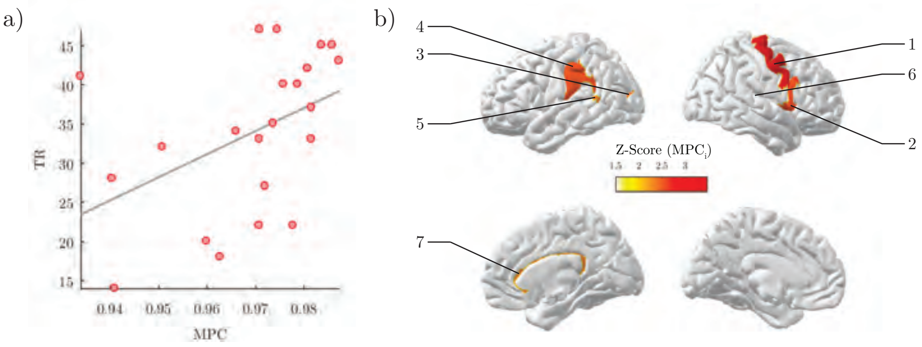


Fig. 2. Network analysis of brain connectivity. a) Scatter plot of the global multi-participation coefficient (MPC) and the total recall (TR) score of AD subjects (Spearman's correlation $R = 0.5547$, $p = 0.0074$). b) Inter-frequency centrality. Statistical brain maps of group differences for local multi-participation coefficients MPC_i. Only significant differences are illustrated ($p < 0.05$, FDR corrected). Z-scores are computed using a non-parametric permutation t-test.

3.1 Methodological considerations

To validate the obtained results we used, in a separate analysis, the imaginary part of coherency [6] to assess functional connectivity. We demonstrated that while no significant between-group differences could be obtained in terms of MPC, the spatial distribution of the MPC values was very similar to that observed in brain networks obtained with the spectral coherence.

4 Conclusion

These findings suggest that multi-layer networks framework reveals complementary information that can be used to identify inter-frequency neural mechanisms of brain diseases and thus give new interpretation possibilities to functional brain connectivity analysis that were not accessible by using more standard - single-layer - approaches.

References

1. Babiloni, C., Ferri, R., Moretti, D.V., Strambi, A., Binetti, G., Dal Forno, G., Ferreri, F., Lanuzza, B., Bonato, C., Nobili, F., Rodriguez, G., Salinari, S., Passero, S., Rocchi, R., Stam, C.J., Rossini, P.M.: Abnormal fronto-parietal coupling of brain rhythms in mild Alzheimer's disease: a multicentric EEG study. Eur. J. Neurosci. 19(9), 2583–2590 (May 2004)
2. Baillet, S., Riera, J.J., Marin, G., Mangin, J.F., Aubert, J., Garnero, L.: Evaluation of inverse methods and head models for EEG source localization using a human skull phantom. Phys Med Biol 46(1), 77–96 (Jan 2001)
3. Battiston, F., Nicosia, V., Latora, V.: Structural measures for multiplex networks. Phys Rev E Stat Nonlin Soft Matter Phys 89(3), 032804 (Mar 2014)
4. Buschke, H.: Cued recall in Amnesia. Journal of Clinical Neuropsychology 6(4), 433–440 (Nov 1984), http://dx.doi.org/10.1080/01688638408401233
5. Grober, E., Buschke, H., Crystal, H., Bang, S., Dresner, R.: Screening for dementia by memory testing. Neurology 38(6), 900–903 (Jun 1988)
6. Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M.: Identifying true brain interaction from EEG data using the imaginary part of coherency. Clinical Neurophysiology 115(10), 2292–2307 (Oct 2004), http://www.clinph-journal.com/article/S1388245704001993/abstract

Bursting synchronization in a neuronal network model for cortical areas of the human brain

Ricardo L. Viana[1] and Fabiano A. S. Ferrari[2]

[1] Universidade Federal do Paraná, Departamento de Física, Curitiba, Paraná, Brazil,
viana@fisica.ufpr.br
[2] Universidade Federal dos Vales do Jequitinhonha e Mucuri, Instituto de Engenharia, Ciência e Tecnologia, Janaúba, Minas Gerais, Brazil,
fabianosferrari@gmail.com

The cerebral cortex can be separated into specialized areas according to specific functions like vision, hearing, olfaction, motor inputs, etc. These cortical areas are connected through a very complex architecture, and the brain is able to integrate information from different cortical areas in order to produce a coherent output. Neuronal synchronization in cortical areas has been observed in mammals [1] and it has been conjectured that synchronization helps to optimize information transfer [2]. In humans, EEG data have revealed connections between neuronal synchronization and higher functions like consciousness and perception [3]. In some cases, however, neuronal synchronization is thought to be related to pathological rhythms and diseases like Parkinson's disease [4]. Accordingly, suppressing such synchronized behavior can be a basis to treatments of Parkinson's disease [5]. Deep brain stimulation is a technique widely used by neurosurgeons for treating Parkinson's disease and essential tremors, and is based on the application of electrical impulses in target areas like the thalamus, subthalamic nucleus and *globus pallidus*, which interfere with and block electrical signals that cause the pathological symptoms [6].

In the present work we propose a mathematical modelling of the suppression of the abnormal neuronal synchronization using a complex network simulating some aspects of the human brain connection architecture provided by Lo *et al.* [7]. We build a network of networks model: in the highest level of description, each node represents a cortical area and the edges represent the density of axonal fibers among cortical areas, determined through diffusion MRI tractography methods. In the lowest level of description, each cortical area is itself a network of coupled individual neurons undergoing a local dynamical behavior, and whose edges are synaptic connections exhibiting small-world properties. The latter hypothesis is based on experimental evidences that cortical area networks in the human brain has such properties [8, 9].

The local dynamics of a neuron is supposed to exhibit spiking activity in two time scales: a fast scale describing autonomous chaotic spiking and a slow scale related to the bursting modulation of the action potential spikes. The essentials of such behavior can be captured by a simple discrete-time model proposed by Rulkov [10]:

$$x_{n+1} = f(x_n, y_n) \equiv \frac{\alpha}{1+x_n^2} + y_n, \qquad y_{n+1} = y_n - \sigma(x_n - \rho),$$

where (x_n, y_n) represent the dynamical variables of the fast and slow scales, respectively, at discrete time $n = 0, 1, 2, \ldots$. The parameters α, σ and ρ are chosen so as to yield autonomous bursting oscillations.

The two-level network description of the neuronal network is represented by

$$x_{n+1}^{(i,p)} = f(x_n^{(i,p)}, y_n^{(i,p)}) + \frac{\varepsilon_e}{2}\left(x_n^{(i-1,p)} - 2x_n^{(i,p)} + x_n^{(i+1,p)}\right) -$$
$$\varepsilon_c \sum_{d=1}^{Q} \sum_{f=1}^{P} \left[W_{(d,f),(i,p)} H(x_n^{(d,f)} - \theta)(x_n^{(i,p)} - V_s)\right],$$

where $x_n^{(i,p)}$ denotes the dynamical variable for the ith neuron belonging to the pth cortical area, $H(..)$ denotes the unit step function, and θ, V_s are coupling parameters. The coupling strength ε_e refers to the local connections in a one-dimensional lattice, and the coupling strength ε_c is for the connections in the upper level, where W_{ij} are the elements of a weighted connectivity matrix.

In numerical simulations we have used 100 neurons in each cortical area with small-world connections chosen according the Newman-Watts procedure: the neurons are connected to their nearest neighbors and have 10% probability of non-local connections, or shortcuts. The cortical areas are coupled to each other through a weighted connectivity matrix based on the data from Lo *et al.* [7]. If $W_{ij} = 1,2,3$ then there are respectively $50,100,150$ links between the cortical areas i and j, such that there are 25% inhibitory and 75% excitatory connections, corresponding to V_s equal to 1 and -2, respectively.

We considered non-identical Rulkov bursting neurons with slightly different values of the parameter α, randomly chosen in the interval $[4.1, 4.3]$ according to a uniform probability distribution. For values within this range every neuron exhibits episodes of chaotic bursting, each of them beginning at discrete time n_k. From a known sequence of such episodes it is possible to define a geometrical phase describing bursting from

$$\varphi_n = 2\pi k + 2\pi \frac{n - n_k}{n_{k+1} - n_k}.$$

Two or more neurons are phase synchronized when their phases are equal, meaning that they start bursting simultaneously, irrespective of the spiking behavior (which is usually non-coherent).

One numerical diagnostic tool to characterize bursting synchronization is the Kuramoto order parameter magnitude [11]

$$R_n = \frac{1}{K}\left|\sum_{j=1}^{K} e^{i\varphi_n(j)}\right|,$$

where K is the number of neurons in the considered assembly. Usually we are more interested in the temporal average of the order parameter magnitude $\overline{R}$. If $\overline{R} = 1$ the neurons are fully synchronized (in the slow time scale of bursting behavior), whereas $\overline{R} < 1$ indicates partial synchronization. This measure can be applied to each cortical area or for the network as a whole, depending on the assembly we study. We have investigated the dependence of the order parameter on the coupling strengths for each cortical area as well as for the network. We observed that, for a large enough value of the coupling strength, the cortical areas with smaller degree exhibit less synchronization (the degree strength is $s_i = \sum_j W_{ij}$ for a given area). Moreover, we observed that the number of cortical areas exhibiting synchronization changes with the coupling strength.

In this work we also investigated the suppression of synchronization due to an external delayed feedback control, in order to simulate the conditions in a deep brain stimulation using our neuronal network as a mathematical model. The delayed control is represented by an injected signal given by

$$M_n(\tau, p) = \frac{1}{Q}\sum_{i=1}^{Q} x_{n-\tau}^{(i,p)},$$

for the pth cortical area and a time delay τ. This signal is applied to the fast variable x of all neurons in a given cortical area. The efficiency of supression can be quantified by the factor

$$S = \sqrt{\frac{Var(M)}{Var(M(\tau, p))}},$$

where M is the mean field in the absence of feedback control and Var stands for the variance. A good suppression of synchronized behavior is achieved with $S \gg 1$. We observed that the network response to this control presents regions of effective suppression as we vary the coupling strength and time delay. When the feedback signal is applied in one subnetwork, for example, no such regions are observed. We have varied the fraction of perturbed subnetworks from 0.25 to 1.0, and the relative size of the suppression regions decreases with this fraction, although the value of the efficiency factor becomes higher.

References

1. R. D. Traub, M. A. Whittington, I. M. Stanford and J. G. R. Jeffreys, Nature **383** (1996) 621.
2. A. Buehlmann and G. Deco, PLOS Computational Biology **6** (2010) e100934
3. L. Melloni, C. Molina, M. Pen, D. Torres, W. Singer, and E. Rodriguez, J. Neuroscience **27** (2007) 2858.
4. P. Silberstein *et al.*, Brain **128** (2005) 1277.
5. C. Hammond, H. Bergman, and P. Brown, Trends in Neurosciences **30** (2007) 357.
6. D. S. Andres, F. Gomez, F. A. S. Ferrari, D. Cerquetti, M. Merello, R. L. Viana, and R. Stoop, Phys. Rev. E **90** (2014) 062709
7. C.-Y. Lo *et al.*, J. Neuroscience **15** (2010) 16876
8. Y. He, Z. J. Chen, and A. C. Evans, Cerebral Cortex **17** (2007) 2407.
9. C. J. Stam and E. C. W. van Straten, Clinical Neurophysiology **123** (2012) 1067.
10. N. F. Rulkov, Phys. Rev. Lett. **86** (2001) 183
11. C. A. S. Batista, S. R. Lopes, R. L. Viana, and A. M. Batista, Neural Networks **23** (2010) 114

Part IV

Diffusion and Epidemics

Heuristic Algorithms for Influence Maximisation in Partially Observable Social Networks

Soheil Eshghi[2], Setareh Magshudi[2,3], Valerio Restocchi[1], Leandros Tassulias[2], Rachel K. E. Bellamy[4], Nicholas R. Jennings[5], Sebastian Stein[1]

[1] University of Southampton, Southampton, UK)
[2] Yale University, New Haven, USA
[3] Technische Universität Berlin
[4] IBM T.J. Watson Research, Yorktown Heights, NY, USA
[5] Imperial College London, UK

1 Summary

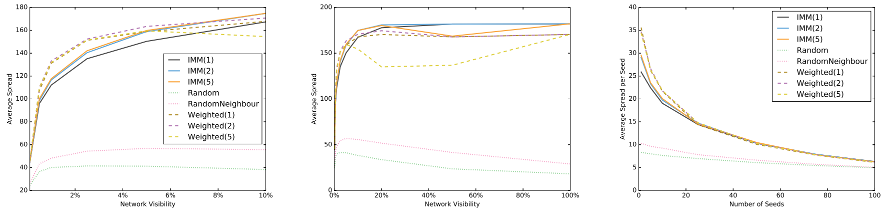
Models to study the propagation of opinions and influence in social networks have been extensively studied and, in particular, the computer science literature has focused on developing algorithms to maximise the spread of influence. However, little work considers the common real-world scenario in which only portions of the full network are visible or only a subset of nodes can be chosen to spread influence from. In particular, in this paper we explore influence maximisation under a type of uncertainty which has not been investigated so far. In our setting, a part (or some parts) of a network is known (e.g., individuals that belong to the decision maker's organisation), while the rest is completely unobservable.

We propose a set of heuristic algorithms designed to maximise the spread of influence in such a setting, by preferentially targeting boundary nodes. We consider the case of organisation-partitioned networks, i.e., networks in which a subset of nodes (a community) and all links among them are fully visible, but the rest are unknown. We show that, in such a setting, the proposed algorithms outperform the state of the art by up to 38%.

2 Method

We propose three heuristic algorithms to select seed nodes:

- **Random Selection:** In this case, we select the seeds simply at random.
- **Random with Neighbour Activation:** Here, we first select the seeds at random, and then for each seed, we activate one of its neighbours. This approach is based on the friendship paradox [2]. It states that on an average basis, most people have fewer friends than their friends have. Thus, if we select some seed at random, it is beneficial to activate one of its neighbours, instead of the original node itself, due to possible larger number of connections.
- **Selection based on (Weighted) Degree:** In this heuristic, we first rank the nodes in the known part of the network based on their degree, so that nodes with larger



(a) (b) (c)

Fig. 1: Figures a and b show the average spread in partially observable networks for 5 seeds. Figure c shows the average spread for varying numbers of seed with visibility 1%.

number of neighbours have higher ranks. Then we select node with highest rank as seeds. Here we use the intuition that nodes with larger number of connections are very likely to be highly influential. In the weighted version of this approach, we still rank the nodes based on their degree; however, in order to improve the influence probability in the unknown part of the network, we attach a higher weight for neighbours that are in the boundary set.

To compute the propagation of influence, we use the NetHept dataset, a network of 15k nodes and 31k edges (representing citations within the high energy physics theory community). We choose this because it constitutes a real-world dataset of reasonable size and because it has been widely used to benchmark influence maximisation algorithms [1]. We compare the performance of the proposed heuristics (measured as the average number of nodes influenced per seed) with that of the most successful influence maximisation algorithm with theoretical performance guarantees (see [3]), namely IMM [1].

3 Results

Figures 1a and 1b show the average spread of all algorithms in a setting with five seed nodes, as we vary the network visibility, i.e., the proportion of fully observable nodes. As expected, the state-of-the-art IMM algorithm performs will throughout all settings. However, looking more closely at cases with low observability (Figure 1a), some of the heuristic approaches outperform it. Specifically, the degree-based heuristics perform consistently well, sometimes achieving an up to 19% higher average spread than IMM. As visibility rises (also continuing in Figure 1b), this difference becomes less pronounced, and by 10% visibility, they achieve a similar performance. As visibility rises further, IMM(1) eventually achieves the highest performance (from 50% visibility onwards).

Looking specifically at the effect of adding higher weights for boundary nodes to the heuristics (denoted as WD(w) and IMM(w), where w is the weight attached to boundary nodes), this can lead to a significant increase in the average spread. However, the performance is sensitive to the exact parameter value, and for networks with higher visibility, a high weight can indeed lead to a decrease in performance, and is most pronounced for Weighted(5) in settings with 20-50% visibility. This is because the boundary nodes actually decrease in importance in the network as more of it is known to the algorithm.

Finally, Figure 1c shows the average spread per initial seed chosen as the number of initial seeds is increased (in a setting with 1% visibility). This highlights that our heuristic techniques achieve the highest performance gains over the state of the art in settings with fewer initial seeds (specifically, a gain of up to 38% when there is just a single initial seed).

Overall, these are promising results, showing that in settings where large parts of the network are not observable and where only few seeds can be chosen, the state-of-the-art algorithm does not necessarily perform best. Instead, simple heuristics perform well, and both those heuristics and the current state of the art benefit from explicitly favouring nodes at the boundary of the known network. It should also be noted that the heuristics are several order of magnitude faster than IMM —a typical run of IMM took about 0.1-0.2 seconds, while the degree-based heuristics typically completed within 0.2-0.3ms.

The aim of this work is to show that current algorithms to maximise influence do not perform well under partial network observability, and are outperformed even by simple heuristics. However, in future work we intend to approach the problem of influence maximisation with partial observability in a more principled way, merging techniques from computer science and statistical physics (see, for example, [4]).

Acknowledgement

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copy-right notation hereon. The work of S. M. was supported by a post-doctoral fellowship from the German Research Foundation (DFG) under Grant Number MA 7111/1-1.

References

1. Tang, Y. and Xiao, X. and Shi, Y.: Influence Maximization: Near-optimal Time Complexity Meets Practical Efficiency. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (2014)
2. Zuckerman, E. and Jost, J.T.: What makes you think you're so popular? Self-evaluation maintenance and the subjective side of the "friendship paradox". Social Psychology Quarterly (2001)
3. Arora, A., and Galhotra, S., and Ranu, S.: Debunking the myths of influence maximiza- tion: An in-depth benchmarking study. Proceedings of the 2017 ACM International Conference on Management of Data (2017)
4. Morone, F., and Min, B., and Bo, L., and Mari, R. and Makse, H: Collective Influence Algorithm to find influencers via optimal percolation in massively. Scientific Reports (2016)

The Cooperation Evolution with Strategy Memory Span in the Weighted Duplex Networks

Jianyong Yu[1*], J.C.Jiang[2], Leijun Xiang[3]

[1] School of Computer Science and Engineering, Hunan University of Science and Technology,
Xiangtan 411201, China.
yujyong@hnust.edu.cn
[2] School of Computer Engineering, Nanyang Technological University, Singapore 639798,
Singapore
[3] College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China

1 Introduction

The diversity of large-scale information, for example, epidemics, gossips, innovation and opinions, spread on social networks. To the understanding of information diffusion in social networks, some threshold and cascade models have been applied [1],[2]. Specially, strategy interactions and diffusion on multiplex networks are generally modeled as evolutionary game [3], which helps to understand the cooperative behaviors and mechanism between selfish individuals as well as sub-networks [4].

In our previous work, the strategy interactions and diffusion between and within communities in the weighted multiplex networks were studied [5]. In this work, we will further study the effect of individuals' trust on cooperation evolution. In fact, the different levels of trust on acquaintances do play a role in the information spreading. Here, the levels of trust correspond to the numbers of consecutive identical cooperation strategies in the strategy memory span of individuals. The results show that the memory of previous strategies highly affects the cooperation evolution. The different strategy memory span is also the main influence factor for the information diffusion, together with the interlayer interaction tight degrees between duplex networks.

2 Method

In order to imitate real networks as far as possible, the heterogeneity of population is considered here. In the intralayer network, the weighted value of linked edges and the link degree of each individual are generated randomly, which represent individual's social position and influence force. In the interlayer network, individuals' neighbors in opposite network are also randomly assigned. Each network is divided into two groups, authority group and non-authority group. The payoffs of an individual are obtained from intralayer and interlayer interactions of strategy, through playing the Prisoner's Dilemma (PD, $1\leqslant T\leqslant 2$ and $-1\leqslant S\leqslant 0$, R=1 and P=0) and the Snowdrift game (SG, $1\leqslant T\leqslant 2$ and $0\leqslant S\leqslant 1$, R=1 and P=0). There are two kinds of individuals adopting different strategies in the games, cooperators (C) and defectors (D). Each individual

adfa, p. 1, 2011.

has a certain strategy memory span, which storage previous strategy status. The ranks of each strategy in different groups are calculated after each interaction. The impact forces of strategy from the different groups to individuals and the reactive forces from individuals are simultaneously taken into account in intralayer and interlayer interactions. An individual will adopt the strategy of authority group in the same layer, or adopt the strategy of a neighbor from the opposite layer with probabilistic parameter a, which reflex the interaction tight degree between two networks. At the stage of updating strategy of individual, previous strategy values in strategy memory spans stand for levels of trust on information diffusion. When consecutive identical cooperation strategies occur in the certain strategy memory span of individuals, the strategy C will be adopted at next intralayer and interlayer interactions.

3 Results

For each pair of game parameters T and S in their value ranges, 1000 weighted duplex networks $C1$, $C2$ with 1000 individuals, and interlayer networks are generated randomly. For each set of $C1$, $C2$ and their interlayer network, the intralayer and interlayer strategy interactions and updates are executed for 5000 times. Lastly, the average densities of cooperators of 1000 random multiplex networks for each pair of T and S are calculated, corresponding to different strategy memory span (ML). Fig.1 shows that the effects of strategy ML on the cooperation diffusion in game SG. With increase of the strategy ML, the density of cooperators C decreases significantly in some regions, especially in regions of $T \in [1.5,2]$ and $S \in [0,0.3]$.

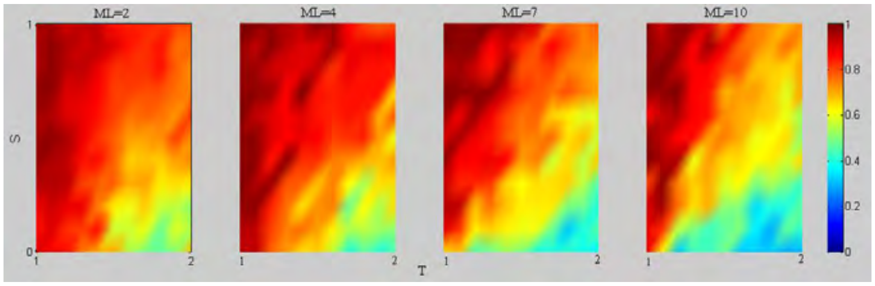


Fig. 1. The average density distribution of cooperators C of 1000 random weighted duplex networks for each pair of T and S in game SG, when the parameter a is 0.9 and strategy ML takes different values.

Moreover, we also consider that the combined effects of parameter a and strategy ML on the density of cooperators C. From the Fig.2, it can be found that the bigger interaction tight degree between two networks is, the higher the density of C is. Meanwhile, the smaller strategy ML is, the higher the density of C is. Specially, the density of C reaches a maximum value when strategy ML is 2. It is easy to be understand that only if one pervious strategy value is C, the next strategy will continue to adopt *C*. This is an important role of the trust levels on cooperation diffusion and evolution. Obviously, it is memoryless case when strategy ML is 1.

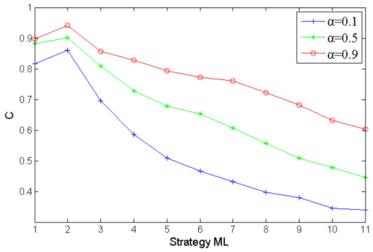


Fig. 2. The impacts of the strategy memory span on densities of *C* in a case study (*T*=1.2 and *S*=0.2 for game SG), and the parameter *a* are 0.9 (red), 0.5(green), 0.1(blue), respectively.

4 Conclusion

In this work, we consider some real scenario about cooperation diffusion in social networks. Some factors affecting individual behavior are taken into account: the social position of individuals, the interaction tight degree between two networks, and the trust levels of individuals on acquaintances by using the strategy memory span. These results provide clues to understand the effects of trust levels of individuals and interaction tight degree on cooperation evolution in the weighted duplex networks.

Acknowledgements

This work was supported by the Natural Science Foundation of China Hunan Province (No.2016JJ2057); the Science Foundation of China Hunan Provincial Education Department (No.15C0546); the Science and Technology Project of Education Department of Fujian Province (No. JA15029).

References

1. S.V.Buldyrev, R.Parshani, G.Paul, H.E.Stanley & S.Havlin, Catastrophic cascade of failures in interdependent networks, Nature 464, 1025–1028 (2010)
2. Z.Wang, L.Wang, A.Szolnoki & M.Perc, Evolutionary games on multilayer networks: A colloquium, Eur. Phys. J. B 88, 124 (2015)
3. M.D.Santos, S.N.Dorogovtsev & J.F.F.Mendes, Biased imitation in coupled evolutionary games in interdependent networks, Sci. Rep. 4, 4436 (2014)
4. H.Lugo & M.San Miguel, Learning and coordinating in a multilayer network, Sci. Rep. 5, 7776 (2015)
5. Jianyong Yu, J.C.Jiang, Leijun Xiang, Group-based strategy diffusion in multiplex networks with weighted values, Physica A: Statistical Mechanics and its Applications, Vol.469, 148-156 (2017)

Targeted immunization in networks with non-overlapping community structure

Zakariya Ghalmane[1], Mohammed El Hassouni[1] and Hocine Cherifi[2]

[1] LRIT, Associated Unit to CNRST (URAC No 29) - Faculty of Sciences, Mohammed V University, B.P.1014 RP, Rabat, Morocco,
[2] LE2I UMR 6306 CNRS, University of Burgundy, Dijon, France

1 Introduction

Finding efficient ways to control the spreading of an epidemic is a challenging issue in a variety of fields, such as medicine, opinion dynamics and computer networks. Research has been conducted on targeted immunization strategies, which make use of the dynamical properties of the infection as well as topological properties of the network structure. However, although most real-world networks exhibit a community structure, few works have been devoted to investigate the impact of this property on epidemic spreading [1-3]. In order to exploit more efficiently the topological properties of real-world networks, we propose and investigate three deterministic immunization strategies based on the network community structure. In this work we restrict our attention to non-overlapping community structure, i.e, a node belongs to a single community. The proposed strategies exploit information such as the size of communities, the number of communities attached to each node and the interconnection density between communities in order to characterize and to rank the influential nodes. Numerical simulations with the Susceptible-Infected-Removed (SIR) epidemiological model are conducted on both real-world and controlled synthetic networks. The effectiveness of the proposed immunization strategies are compared to classical alternatives. Results demonstrate that using information on the community structure is of great importance in order to control disease dynamics.

2 Targeted immunization scheme

The goal of targeted immunization is to change the structure of the network of susceptible individuals so that it is harder for a pathogen to spread. According to a given immunization algorithm, a set of nodes is chosen and their state is set to resistant. Deterministic algorithms rank the nodes according to a measure of their ability to limit the propagation process. Target nodes are then chosen by their rank from high to low, until a desired immunization coverage of the population is achieved. Indeed, the number of people to remove from the susceptible class is often constrained. Therefore the problem translates into ranking the nodes according to a centrality or influence value. This centrality must be related to features of both the propagation process and the network structure. Note that if two or more nodes share the same centrality values their rank is assigned at random. To investigate the influence of the community structure in the propagation process, we propose three measures that integrate various levels of information.

Number of Neighboring Communities: The main idea of this measure is to rank nodes according to the number of communities they reach directly (through one link). The reason for targeting these nodes is that they are more likely to contribute to the epidemic outbreak towards multiple communities. Note that all the nodes that do not have inter-community links share the same null value for this measure.
For a given node i belonging to a community $C_k \subset C$, it is given by:

$$m_i = \sum_{C_l \subset C \setminus \{C_k\}} \bigvee_{j \in C_l} a_{ij} \tag{1}$$

Where a_{ij} is equal to 1 when a link between nodes i and j exists, and zero otherwise. $\bigvee$ represents the logical operator of disjunction, i.e, $\bigvee_{j \in C_l} a_{ij}$ is equal to 1 when the node i is connected to at least one of the nodes $j \in C_l$.

Community Hub-Bridge measure: Each node of the network share its links with nodes inside its community (intra-community links) and nodes outside its community (inter-community links). Depending of the distribution of these links, it can propagates the epidemic more or less in its community or to its neighboring communities. Therefore, it can be considered as a hub in its community and a bridge with its neighbors communities. That is the reason why we call this measure the Community Hub-Bridge measure. Furthermore, the hub influence depends on the size of the community, while the bridge influence depends on the number of of its neighbors communities. For a given node i belonging to a community $C_k \subset C$, it is given by:

$$l_i(C_k) = h_i(C_k) + b_i(C_k) \tag{2}$$

With $h_i(C_k) = Card(C_k) * k_i^{intra}(C_k)$ and $b_i(C_k) = m_i * k_i^{inter}(C_k)$.
Where $k_i^{intra}(C_k)$ and $k_i^{inter}(C_k)$ are respectively the intra-community degree and the inter-community degree of the node i. $Card(C_k)$ is the size of its community. m_i represents the number of its neighboring communities that are connected to the node i.
$h_i(C_k)$ tend to immunize preferentially hubs inside large communities. Indeed they can infect more nodes than those belonging to small communities. $b_i(C_k)$ allows to target nodes that have more links with various communities.
Community Hub-Bridge with link Density measure: for a given node i belonging to a community $C_k \subset C$, it is given by:

$$d_i(C_k) = \rho_{C_k} * h_i(C_k) + (1 - \rho_{C_k}) * b_i(C) \tag{3}$$

Where ρ_{C_k} represents the interconnection density between the community C_k and the other communities of the network. It is given by:

$$\rho_{C_k} = \frac{\sum\limits_{i \in C_k} k_i^{inter} / (k_i^{inter} + k_i^{intra})}{Card(C_k)} \tag{4}$$

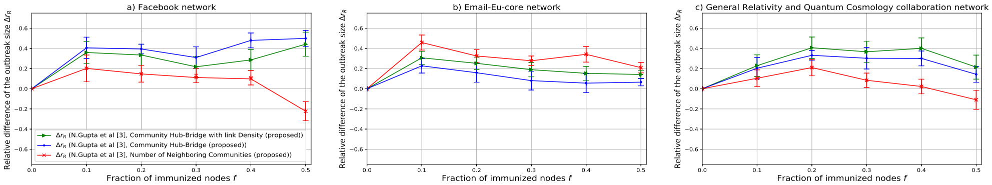


Fig. 1. The relative difference of outbreak size Δr_R between Comm measure and the proposed measures, performed on the real-world networks.

If the communities are very cohesive then more importance is given to the bridges in order to isolate the communities. Otherwise, more importance is given to the hubs.

3 Results

We performed a series of experiments on both real-world and synthetic networks with a known community structure in order to investigate the efficiency of the proposed measures. Here we report the results of the comparative evaluation with the Comm measure proposed by Naveen et al [3]. Indeed, it is the most efficient alternative measure. The classical SIR model is used to investigate the spread of epidemics. To evaluate the effectiveness of the proposed measures, we opt for the relative difference of outbreak size. It is defined as $\Delta r_R = (R_{ref} \text{ \_ } R_p)/R_{ref}$, where R_{ref} and R_p are respectively the final number of recovered nodes for the reference and proposed measure. If Δr_R is positive, the epidemic spreads less when applying the proposed measure.

In Fig.1 we evaluate our approach on three network datasets: ego-Facebook[3], Email-Eu-core[3] and ca-GrQc[3] networks. It is obvious that the Number of Neighboring Communities measure outperforms the Comm measure. It does not, however, target community hubs. That is the reason why it performs worse than Comm measure in the case of a large immunization coverage as illustrated in Fig.1 a) and c) when the fraction of immunized nodes f is more than 0.5. We note also from Fig.1 that the other proposed measures outperform the Comm measure for the different values of *f*.

To summarize, the proposed algorithms perform better as they are able to extract more useful information on community structure. Experimental results reveal that they are very effective in identifying the influential nodes.

References

1. M. Salathé, J. H. Jones, (2010). Dynamics and control of diseases in networks with community structure. PLoS computational biology, 6(4), e1000736.
2. C. Stegehuis, R. van der Hofstad & J. S. van Leeuwaarden, (2016). Epidemic spreading on complex networks with community structures. Scientific reports, 6.
3. N. Gupta, A. Singh, H. Cherifi, (2016). Centrality measures for networks with community structure. Physica A: Statistical Mechanics and its Applications, 452, 46-59.

[3] http://snap.stanford.edu/data/

Modeling Influence Diffusion in Social Networks for Viral Marketing

Wenjun Wang and W. Nick Street

Department of Management Sciences, University of Iowa, Iowa City, IA 52242, USA
wenjun-wang@uiowa.edu
nick-street@uiowa.edu

1 Introduction

Modeling influence diffusion in social networks is an important challenge. There are a large number of models in the literature addressing influence diffusion and viral marketing [2], [4]. However, there exist some significant limitations in these models. First, they approach a social entity's adoption likelihood from a confined scope, considering only the direct influence from the activated neighbors of the entity. Second, most models overlook the fact that influence does not remain static or constant, but rather attenuates along diffusion paths and decays with time. Third, these models fail to capture the individual temporal diffusion dynamics. In this paper, we propose a novel *multiple-path asynchronous threshold* (MAT) model to address these issues, and develop an effective and efficient heuristic to tackle the influence-maximization problem.

2 MAT Model

We categorize all the nodes in the network into two types: influencers and messengers. An *influencer* is an active node that has adopted the product and can originate and spread its influence in the network. A *messenger* is an inactive node that may acquire influence and pass the influence on to others. Once a messenger acquires influence that is greater than or equal to its threshold, it is activated and turns into an influencer who starts to spread out its own influence. Our model captures an important characteristic of word-of-mouth (WOM) communication in that: *anyone can pass along WOM messages and potentially influence the recipient*. In other words, a node can be activated by not only the *direct influence* from active neighbors (influencers), but also the *indirect influence* passed along by its inactive neighbors (messengers). This is a distinguishing and more realistic feature built in our MAT model.

To differentiate the relationship strength on influence, we introduce a *weight normalization scheme* that measures the fraction of influence a node receives from a specific in-neighbor relative to the total influence it receives from all of its in-neighbors. To quantify the influence attenuation along a diffusion path, we define a depth-associated attenuation coefficient $\alpha = d^{-2}$, where d is the depth (number of hops) from an influencer to the node of interest along the diffusion path. It can be interpreted as a compounding factor that incorporates the trustworthiness decay, information corruption, and decreasing reaching probability. Specifically, we set the depth limit d_{max} to 3 for each influencer to capture the *three-degrees-of-influence* phenomenon [3]. On the other hand, we model the temporal influence decay as an exponential function of time, $I(t) = e^{-\lambda t}$, where λ is a user-specified tunable parameter of decay rate. It can be tuned to account for different products on various social networks. To capture the individual temporal diffusion dynamics, we model the heterogeneity of WOM messaging from a node to its neighbors as a *Poisson process* with a rate that is determined by the node's relative activeness at both local and global level in terms of its out-link weights.

The diffusion process starts with an initial set of influencers (seed nodes) S_0 with $|S_0| = K$, and unfolds in discrete time steps. At each time step, the influence is propagated one hop from a node u to each out-neighbor v with a probability based on their contact frequency and node u's global activeness. Each inactive node v is assigned with an activation threshold selected uniformly at random in the range $[0,1]$. When the total influence that v receives is greater than or equal to it threshold, it is activated and turns into an influencer. Then it not only continues passing other influencers' influence as a messenger, but also starts to spread out its own influence as an influencer. The diffusion process stops when the number of hops of influence diffusion of each influencer reaches the depth limit (set to 3 by default) and no new activation is possible.

3 IV-Greedy Algorithm

The influence-maximization (IM) problem is to find a small set of K seed nodes (initial adopters) who can trigger the largest further adoptions in the network. The IM problem is NP-hard under the MAT model. Using the *influence vector* (IV) of each node, we develop a heuristic algorithm called *IV-Greedy*. The influence vector of a node captures where and how much influence it spreads out in its neighborhood based on a *static* version of the MAT model, in which we ignore the temporal diffusion decay, individual diffusion dynamics, and the activation of any nodes. We use it as a proxy for the *dynamic* influence diffusion so as to avoid the expensive Monte Carlo (MC) simulation. We sweep over the influence vector of each node to repeatedly pick the node with the *maximum marginal gain* and add it to the seed set until all K seeds are found.

4 Experiments

To evaluate our MAT model and the performance of IV-Greedy, we conduct experiments on three widely-used real-life network datasets, which include PGP [1], NetHEPT [1], and WikiVote [2]. We compare the performance of IV-Greedy against a set of baseline algorithms in terms of both influence spread and time efficiency. The simplest baseline is to select the seed nodes uniformly at random. The most frequently used is the *degree-centrality* heuristic, in which the seed nodes are chosen in descending order of their out-degrees. The next baseline is the Top-K algorithm, which selects the top K nodes with the largest *individual* influence spread based on MC simulation. As shown in Fig. 1 and Fig. 2, IV-Greedy achieves the largest influence spread, and it is up to three orders of magnitude faster than Top-K. IV-Greedy is the best performing algorithm overall.

[1] http://research.microsoft.com/en-us/people/weic/graphdata.zip

[2] https://snap.stanford.edu/data/wiki-Vote.html

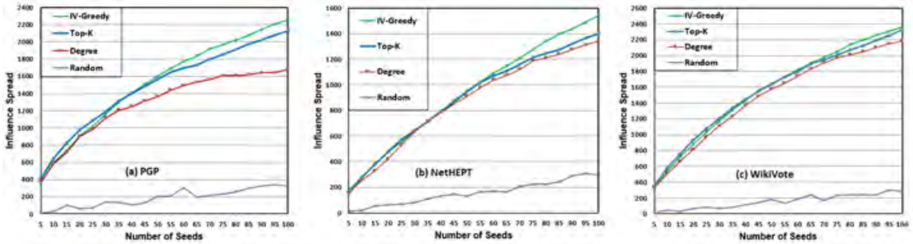


Fig. 1. Performance comparison on influence spread

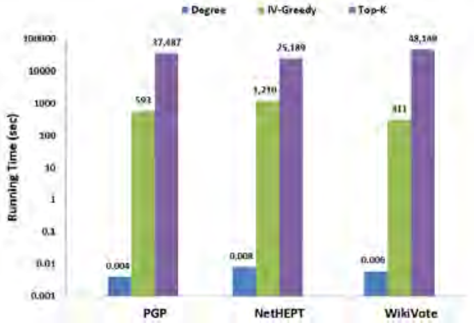


Fig. 2. Performance comparison on running time (CPU seconds)

5 Conclusion

In this paper, we propose a novel *multiple-path asynchronous threshold* (MAT) model for viral marketing in social networks. Our MAT model captures both direct and indirect influence, influence attenuation along diffusion paths, temporal influence decay, and individual diffusion dynamics. It is an important step toward a more realistic diffusion model. Further, we develop an effective and efficient heuristic, IV-Greedy, to tackle the influence-maximization problem. Our experiments on three real-life networks demonstrate its excellent performance in terms of both influence spread and time efficiency. Our work provides preliminary but important insights and implications for diffusion research and marketing practice.

References

1. Boguna, M., Pastor-Satorras, R., Diaz-Guilera, A., Arenas, A.: Models of social networks based on social distance attachment. Phys. Rev. E 70, 056122 (2004)
2. Chen, W., Lakshmanan, L., Castillo, C.: Information and Influence Propagation in Social Networks. Morgan & Claypool Publishers (2013)
3. Christakis, N.A., Fowler, J.H.: The spread of obesity in a large social network over 32 years. New England J of Medicine 357, 370–379 (2007)
4. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. Theory of Computing 11(4), 105–147 (2015)

The importance of the temporal sequence of indirect contacts: the case study of a dairy farm system in the Emilia Romagna region (Northern Italy)

Alba Bernini[1,2], Luca Bolzoni[2], and Renato Casagrandi[1]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, 20133 Milan, Italy
[2] Risk Analysis Unit, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna, 43126 Parma, Italy
alba.bernini@polimi.it
luca.bolzoni@izsler.it
renato.casagrandi@polimi.it

1 Introduction

Among the many real-world processes that can be represented as dynamical systems on networks, epidemics spread is one of the most notable examples. In the case of livestock diseases, the diffusion of epidemics in farm systems can cause serious negative impacts both from economic and social perspectives [**?**]. Thus, quantitative epidemiological studies are key in supporting the design of more effective control measures. In this context, nodes can represent farms, considered as epidemiological units, while links can describe the possible between-farm pathogen transmission routes, which can in turn be distinguished between direct contacts, i.e. animal movements, and indirect ones, such as the sharing of equipment or movement of workers and vehicles. The role of indirect contacts in disease transmission is still poorly understood due to limitations deriving from their highly diverse and complex nature. Indeed, while animal movements (such as bovine and swine) are registered in national databases in many EU-countries (Commission Decision 2006/132/EC), little information is available to date on workers visits. As a consequence, when accounted for in modelling, indirect contacts have been described through the use of commercial networks, where links between farms are traced on the basis of common contractors [**?**]. However, this approach can lead to descriptions of the contact networks that are misleading, since *(i)* a common contractor does not imply common personnel and vehicles visiting a pair of farms and *(ii)* the temporal sequence of contacts is lost. Here, we analyzed how different levels of detail in the representation of indirect contacts may affect the description of the epidemic spread process, to the point that different conclusions can sometimes be obtained. In such cases, the detection of superspreaders, i.e. the farms that play a crucial role in the diffusion process, is falsified and potential control actions might become ineffective.

2 Materials and methods

We considered a system of dairy farms in the Emilia Romagna region (Northern Italy) involved in a comprehensive data collection campaign on calves transportation occurred between September and November 2014. Based on these data, we reconstructed the daily routes of the trucks belonging to different transportation companies active in the area. As a matter of fact, the contamination of trucks (such as for milk, feed, and live animals transportation) represents one of the main indirect route of between-farm transmission. For the same period and the same farms, information about direct contacts were collected from the Italian National Database for Animal Identification and Registration. The between-farm contacts were represented as a daily temporal multilayer network [**?**,**?**] with two layers: one for direct (i.e. animal movements) and the other for indirect contacts (derived from the sharing of transportation trucks). We derived the indirect contacts network using two different levels of detail. Indeed, our aim was to evaluate to what extent the knowledge of the sequence of on-farm visits is relevant to establish the final size of an epidemic in the unfortunate case of a disease diffusion. On the one hand, we built a bipartite network based on the commercial relations between the farms and the transportation companies and we projected it on the space of the farms to obtain the *common contractors network (CCN)*. On the other hand, taking advantage of the available data on the truck identifiers and the sequence of visits, the *truck itineraries network (TIN)* was assembled assigning a directed link from a given farm to those later visited by the same truck in the same day. It is worth remarking that the TIN is different with respect to the CCN not only because of the introduction of the links directionality, but also because of the different number of links, since some transportation companies own more than one truck. To remark the crucial role played by the topology of the network, we kept the description of the disease diffusion process as simple as possible. Therefore, the system was modeled on both multilayer networks through a boolean Susceptible-Infectious (SI) compartmental model, where the probability of disease transmission was designed according to Bates *et al.* [**?**]. At the beginning of each simulation, all farms were assumed susceptible but one (namely, the epidemic seed). One at a time, each farm was selected as epidemic seed and 100 simulations were performed. At the end of each simulation, the final epidemic size was recorded and the 5% of farms that led to the larger final epidemic sizes were classified as superspreaders and characterized in terms of their topological characteristics.

3 Results and discussion

The distributions of the final epidemic sizes obtained with the two multilayer networks CCN and TIN were significantly different, as shown in Fig.1. Each point on the x-axis represents an epidemic seed and the vertical segments surrounding the median final epidemic sizes on the y-axis represent the ranges out of the 100 simulations performed describing the indirect contacts through the CCN (grey) and TIN (black). Farms are ranked in decreasing order of median final epidemic size predicted either using the CCN or the TIN. As emerging from Fig.1, the use of the CCN systematically resulted in an overestimation of the final epidemic size. Interestingly, we found that the superspreading farms identified through the simulations on the CCN (red dots) rarely match the superspreaders in the TIN (Kendalls coefficient between the two rankings equals 0.41). This result has significant consequences on the surveillance and control of livestock diseases, especially in the case of implementation of risk-based biosecurity measures.

Specifically, if the network structure used to identify the superspreaders is inadequate, the intervention will focus on the wrong farms. Among all the indicators considered to potentially characterize the superspreaders, the best resulted to be the out degree and the outgoing infection chain [**?**], which seem to be good candidates for selecting the farms where starting to implement biosafety measures.

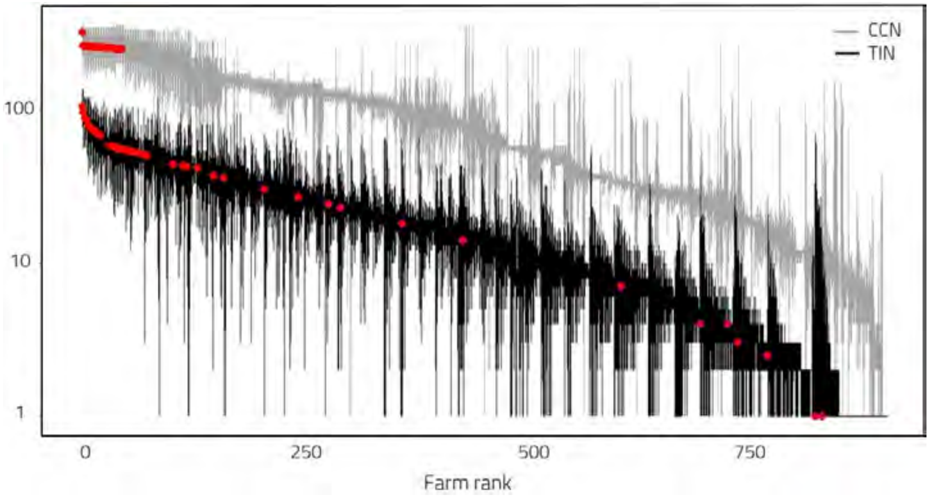


Fig. 1. Comparison between the distributions of the final epidemic size obtained with the two temporal multilayer networks. Each x-axis point represents an epidemic seed and, for each of them, are reported on semi logarithmic scale the 100 final epidemic sizes obtained. The red dots represent the positions in the two rankings of the farms associated to the largest 5% median final epidemic sizes in the commercial network, i.e. the superspreaders.

References

1. Anderson, I.: Foot and mouth disease 2001: lessons to be learned inquiry. The Stationery Office, London, (2002)
2. Fournié, G., Guitian, J., Desvaux, S., Cuong, V. C., Pfeiffer, D. U., Mangtani, P., Ghani, A. C.: Interventions for avian influenza A (H5N1) risk management in live bird market networks. Proc. Natl. Acad. Sci. USA 110, 91779182 (2013)
3. Holme, P., Saramäki, J.: Temporal networks. Physics reports 519(3), 97-125 (2012)
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D. U.: Complex networks: Structure and dynamics. Physics reports, 424(4), 175-308 (2006)
5. Bates, T. W., Thurmond, M. C., Carpenter, T. E.:Description of an epidemic simulation model for use inevaluating strategies to control an outbreak of foot-and-mouth disease. Am. J. Vet. Res. 64, 195204 (2003)
6. Dubé, C., Ribble, C., Kelton, D., McNab, B.: Comparing network analysis measures to determine potential epidemic size of highly contagious exotic diseases in fragmented monthly networks of dairy cattle movements in Ontario, Canada. Transbound. Emerg. Dis. 55, 382-92 (2008)

Partial network immunization in Continuous-Time Information Cascades

Argyris Kalogeratos[1], Kevin Scaman[1*,2], Luca Corinzia[1*,3], and Nicolas Vayatis[1]

[1] CMLA – ENS Cachan, CNRS, Université Paris-Saclay, 94230 Cachan, France
[2] MSR-Inria Joint Center, 91120 Palaiseau, France
[3] ETH Zürich, 8092 Zürich, Switzerland
{kalogeratos, scaman, vayatis}@cmla.ens-cachan.fr, lucac@ethz.ch

1 Introduction

Studying the way in which diffusion processes evolve in networks is fundamental for further understanding such complex and dynamic phenomena. In particular, being able to predict the number of nodes that will be reached at the end of the spread when that starts from a known set of initial infection seeds (i.e. seeds' *influence*) is of broad interest. Taking preparatory measures by acting on the network, so as to reduce the reach of a possible future diffusion, is a core administration problem. Traditionally, that has been studied as a way to improve public healthcare (e.g. through vaccination), recently though it has attracted a lot of attention due to the concerns raised by cases of malicious information propagation in social networks (e.g. fake news, rumors).

In the existing literature the role of the spectral radius of the adjacency matrix representing the underlying network has been largely highlighted as a quantity tightly connected with the epidemic threshold over which the reach of the diffusion explodes and becomes comparable with the network size [8,4,1]. Various studies have been presented for virus propagation models and influence maximization [10,6,5].

In this paper we present a brief overview of our recent work on partial node immunization in Continuous-Time Information Cascade Model ($\mathcal{CTIC}$) [7]. $\mathcal{CTIC}$ [2] is a stochastic model allowing propagation rates along edges to vary in time. Relying on previous work, we use the concept of *Hazard radius* introduced in [4], that is highly correlated to the influence and helps us in deriving upper bounds for the influence under the $\mathcal{CTIC}$. We subsequently develop the *NetShape* strategy that enjoys a convex relaxation and, among other influence optimization tasks that we do not go through in this short summary, it can be used for *offline and partial node immunization*. In that scenario, a budget of treatment units is available. Each treatment unit can target a single node, in advance of the diffusion, thereafter reducing node's propagation rates along all of its outgoing edges by a fixed factor.

2 Results

The NetShape method. Formally, let $\mathcal{F}_{ij}(s-\tau_i)$ be the *Hazard function*, an element of the *Hazard matrix* $\mathcal{F}$, representing the propagation rate on edge $i \rightarrow j$ at a specific time

[*] Part of the work has been conducted while author was at CMLA[1].

s after τ_i when node i received the piece of information and got 'infected'. Also, let the *Hazard radius* be the spectral radius of a matrix computed by integrating the Hazard functions over time (i.e. the component-wise integration of the symmetrized $\mathcal{F}$):

$$\rho_H(\mathcal{F}) = \rho\left(\int_0^{+\infty} \frac{\mathcal{F}(t) + \mathcal{F}(t)^{\mathsf{T}}}{2} dt\right), \tag{1}$$

where $\rho(\cdot) = \max_i |\lambda_i|$, and λ_i are the eigenvalues of the implied input matrix (since we refer to square matrices). By further elaborating results from [4], we have shown that the maximum influence cannot exceed a certain proportion of the network that is non-decreasing with $\rho_H(\mathcal{F})$, and displays a sharp transition between a sub-critical and super-critical regime. Therefore, we solve the following optimization problem over a set of feasible Hazard matrices $\mathbb{F}$ that can be produced by valid actions on the nodes:

$$\mathcal{F}^* = \mathrm{argmin}_{\mathcal{F} \in \mathbb{F}} \ \rho_H(\mathcal{F}). \tag{2}$$

When $\mathbb{F}$ is a convex set, this optimization problem is also convex and the proposed *NetShape* method uses a simple *projected subgradient descent* scheme to solve it. The interested reader is refereed to [7] for more technical details on NetShape algorithm.

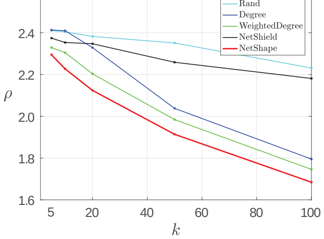
Experimental evaluation. We evaluated the NetShape algorithm for the *offline and partial node immunization* under the $\mathcal{CTIC}$ and compared it with baseline and state-of-the-art approaches for selecting the k nodes to target (k is the provided budget): **i)** random node selection (*Rand*); **ii)** selection of the nodes with the highest out-degree (*Degree*); **iii)** selection of k nodes with highest sum of outgoing edge weight $w_{ij} = \int_0^{+\infty} \mathcal{F}_{ij}(t)dt$ (*WeightedDegree*), actually derived by the optimization of the lower bound LB_1 in [3]; **iv)** the *NetShield* algorithm from [9] (originally designed for total immunization).

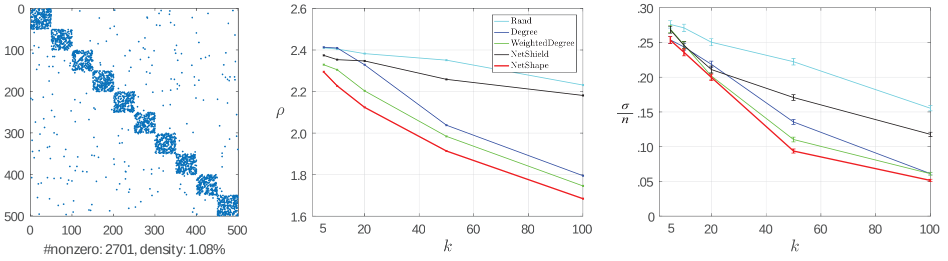
For our empirical evaluation we used an artificial random network with $n = 500$ nodes generated as follows: 10 equally-sized Erdős Rényi clusters were first created with edge creation probability $p = 0.1$, then their adjacency matrices were synthesized in a block-diagonal structure with a uniform inter-cluster rewiring probability $p' = 0.001$. Fig. 1a shows the structure of the adjacency matrix. Finally, the weights of the created edges (i.e. the transmission probabilities) were generated using a *trivalency model* that picks values uniformly at random from the set $\{low\colon 0.1, med\text{ium}\colon 0.2, high\colon 0.5\}$.

Each treatment budget can be assigned to a single node and, here, we assume that it can cause a fixed decrease of 70% in that node's propagation rate along all of its edges. Fig. 1b, c plot the curves (average values and stds) of two evaluation measures for our simulations over a set of budget sizes. For each k value, we run 1000 simulations and each simulation starts from nodes of high influence. The measure reported in Fig. 1 c is the influence of the selected seeds, i.e. the expected proportion $\frac{\sigma}{n}$ of infected nodes at the end of the process. Also, the measure plotted in Fig. 1 c is the spectral radius $\rho_H(\mathcal{F})$ of the Hazard matrix that NetShape minimizes as a proxy for influence reduction.

Note that our purpose was to test in a meaningful parametrization scenario where the spectral radius of the original network would have been close to 1 and, thus, its decrease could cause a non-negligible reduction to the influence.

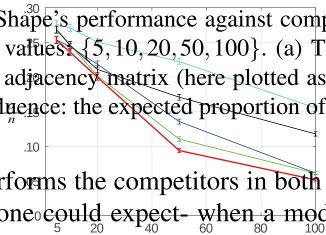
Performance results. The brief reported results show that: **i)** NetShape optimizes the spectral radius $\rho_H(\mathcal{F})$ (an empirical proof of correctness for our optimization scheme), **ii)** effectively minimizes the influence (verifying the relevance of our optimization to





(a) (b) (c)

Fig. 1. Comparison of NetShape's performance against competitors on an artificially generated random network. Tested k values: $\{5, 10, 20, 50, 100\}$. (a) The structure of the generated non-symmetric, block-diagonal adjacency matrix (here plotted as binary matrix); (b) spectral radius $\rho_H(\mathcal{F})$ vs. budget k; (c) influence: the expected proportion of infected nodes $\frac{\sigma}{n}$ vs. budget k.



the influence), **iii)** outperforms the competitors in both previous points; the largest difference is observed -as one could expect- when a moderate amount of treatments are available. In conclusion, the presented approach seems promising and we plan to investigate its potential generalization to other influence optimization problems.

References

1. Chen, C., Tong, H., Prakash, B.A., Tsourakakis, C.E., Eliassi-Rad, T., Faloutsos, C., Chau, D.H.: Node immunization on large graphs: Theory and algorithms. IEEE Transactions on Knowledge and Data Engineering 28(1), 113–126 (2016)
2. Chen, W., Lakshmanan, L.V., Castillo, C.: Information and influence propagation in social networks. Synthesis Lectures on Data Management 5(4), 1–177 (2013)
3. Khim, J.T., Jog, V., Loh, P.L.: Computing and maximizing influence in linear threshold and triggering models. In: Proceedings of the Advances in Neural Information Processing Systems 29. pp. 4538–4546 (2016)
4. Lemonnier, R., Scaman, K., Vayatis, N.: Tight bounds for influence in diffusion networks and application to bond percolation and epidemiology. In: Advances in Neural Information Processing Systems. pp. 846–854 (2014)
5. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 420–429 (2007)
6. Ohsaka, N., Akiba, T., Yoshida, Y., Kawarabayashi, K.i.: Fast and accurate influence maximization on large networks with pruned Monte-Carlo simulations. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 138–144 (2014)
7. Scaman, K., Kalogeratos, A., Corinzia, L., Vayatis, N.: A spectral method for activity shaping in continuous-time information cascades. ArXiv e-prints (Sep 2017)
8. Scaman, K., Lemonnier, R., Vayatis, N.: Anytime influence bounds and the explosive behavior of continuous-time diffusion networks. In: Advances in Neural Information Processing Systems. pp. 2017–2025 (2015)
9. Tong, H., Prakash, B.A., Tsourakakis, C., Eliassi-Rad, T., Faloutsos, C., Chau, D.H.: On the vulnerability of large graphs. In: Proceedings of the IEEE International Conference on Data Mining. pp. 1091–1096 (2010)
10. Wang, Y., Chakrabarti, D., Wang, C., Faloutsos, C.: Epidemic spreading in real networks: An eigenvalue viewpoint. In: Proceedings of the IEEE International Symposium on Reliable Distributed Systems. pp. 25–34 (2003)

Epidemic Conductance in complex networks

Joan T. Matamalas[1], Alex Arenas[1], and Sergi Gómez[1]

Departament d'Enginyeria Informtica i Matemtiques
Universitat Rovira i Virgili
43007 Tarragona, Spain

The problem of modeling the spread of a disease among individuals has been studied in deep over many years. The development of compartmental models, that divide the individuals among a set of possible states, has given rise to a new collection of technics that enables, for instance, the analysis of the epidemic threshold or the study of the impact of a prophylactic campaign. After the initial epidemiological studies on well-mixed populations, it has been recognized that complex networks constitute a better description for the substrate on top of which the epidemic spreading takes place. Among the many available epidemic models, the Susceptible-Infected- Susceptible (SIS) has become a cornerstone in the study of epidemic spreading in complex networks. From the initial analysis of SIS using heterogeneous mean field approximations to determine the epidemic threshold [1], to the recent ones in which the probability of being infected is determined at the level of node [2], there have been uncountable advances on this topic [3].

In this work we analyze the SIS model in complex networks at the level of edges. In particular, we propose the definition of the epidemic conductance as the probability that a link is in condition of spreading the epidemics. We show how to obtain equations for the conductance of all the links, which can be solved by iteration in a similar way to the Microscopic Markov Chain Approach (MMCA) in [2]. These equations provide a more accurate description of the global epidemic incidence and the epidemic threshold than previous methodologies.

The idea of analysing the epidemic process at a level of links is simple, but the consequences are enormous. For example, if we identify the links which are more involved in the propagation of a disease, it is possible to design targeted countermeasures which affect just specific links instead of whole nodes, while being more effective, Fig. 1. This can be illustrated by a hypothetical pandemic disease propagated using the air transportation network: the isolation of one airport is a dramatic measure that is socially and politically difficult to accept and put into practice, but the suspension of just a few connections between selected airports could be more easily assumed, and at the same time achieving a better contention of the disease.

To summarize, the aim of our work is to show the valuable information which can be extracted when we analyze the contribution of the links to dynamical processes in complex networks. In particular, we are able to fully describe an epidemic spreading using equations for all the links, where the variables describe important magnitudes of the dynamics, such as the conductance of the links, i.e. the probability that a link is in a configuration which enables the spreading of the disease. We also show how removing the links with largest conductance enhance the containment of the disease, more efficiently than acting on nodes, as in the previous airports example. Similarly, we also indicate how to deal with other spreading processes, initiating an innovative approach in the analysis of dynamics in complex networks.

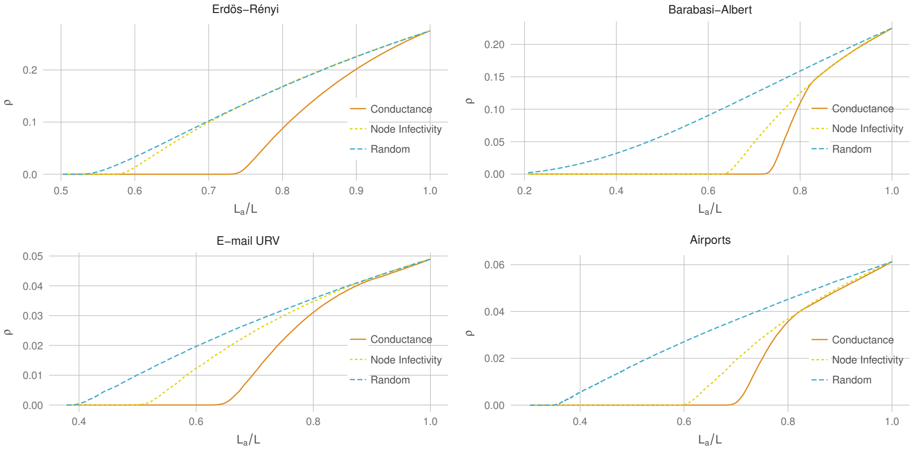


Fig. 1. Targeted edge percolation. We show the incidence of the epidemics, ρ, as function of the occupation probability, L_a/L, where L_a is the current number of active edges in the percolation process. We compare three different percolation strategies: a random edge removal (blue dashed line); removing the edges of the node with highest probability of being infected, $P(\sigma_i = I)$ (yellow dotted line); and removing the edge that has the largest total conductance (orange solid line). We have made use of the total conductance since the percolation process is undirected and we try to remove as much conductance as possible

References

1. Pastor-Satorras, R. and Vespignani, A.: Epidemic spreading in scale-free networks. Physical Review Letters, 86 (2001) 3200.
2. Gomez, S., Arenas, A., Borge-Holthoefer, J., Meloni, S. and Moreno, Y.: Discrete-time Markov chain approach to contact-based disease spreading in complex networks. EPL (Europhysics Letters), 89 (2010) 38009.
3. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. and Vespignani, A.: Epidemic processes in complex networks. Reviews of Modern Physics, 87 (2015) 925.
4. Matamalas, J. T., Arenas, A. and Gomez, S.: Epidemic conductance in complex networks. Preprint (2017) in preparation.

Synergistic cumulative contagion in epidemic spreading

Xavier R. Hoffmann[1,2] and Marián Boguñá[1,2]

[1] Departament de Física de la Matèria Condensada, Universitat de Barelona, Spain
[2] Universitat de Barcelona Institute of Complex Systems (UBICS), Spain

hoffmann@ub.edu
marian.boguna@ub.edu

1 Introduction

Epidemic modeling has proven to be a powerful tool for the study of spreading and contagion phenomena in biological, social and technical systems. The addition of numerous compartments and the incorporation of complex contact topologies has yielded evermore accurate models, prompting their use as real-time predictive tools [5]. Notwithstanding, most approaches assume memoryless, isolated and independent processes, an approximation partially invalidated by empirical evidence [1, 2]. We propose an alternative synergistic and cumulative infection mechanism, and study its effects in the susceptible-infected-susceptible model [3, 4].

2 Methods

In our description, susceptible agents accumulate pathogens from all their infected neighbors and become infected following a given probability density. When the last infected neighbour of a susceptible agent recovers, its accumulated viral load starts to decay with a characteristic relaxation time ζ. Infected agents recover spontaneously following a given inter-event time distribution.

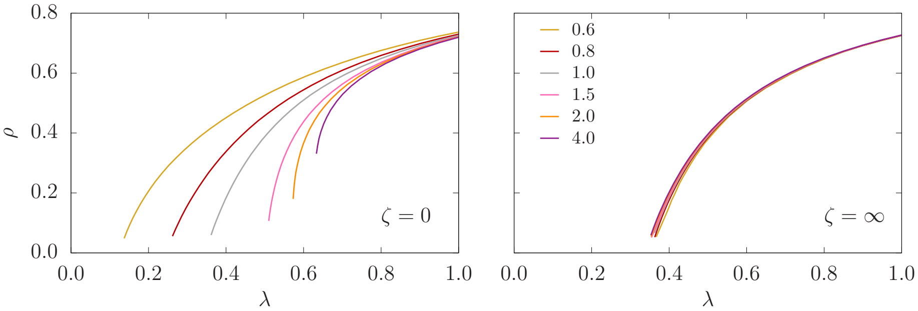


Fig. 1. Fraction of infected agents, ρ, as a functon of the effective spreading ratio, λ, for various infection distributions (α indicated in legend). **Left**: Instantaneous decay of viral load (short-term memory). **Right**: Viral load does not decay (long-term memory).

Here we use a Weibull distribution for infections (with shape parameter α) and exponentially distributed recoveries. We study the limit cases $\zeta = 0$ (instantaneous decay) and $\zeta = \infty$ (perpetual accumulation) in random degree regular networks. We characterize the system's stationary and dynamical properties by means of extensive numerical simulations. In particular, we analyze the differences in approaching the steady state from above (starting from a fully infected population) and below (starting with a single infected agent).

3 Results

When agents solely present short-term memory ($\zeta = 0$) the critical point of the transition between the healthy and endemic phases varies greatly depending on the shape of the infection probability density (left panel Fig. 1). Moreover, and even though the phase transition remains continuous, mean-field universality is lost.

If individuals are equipped with long-term memory ($\zeta = \infty$), the system experiences a rather counterintuitive collective memory loss (right panel Fig. 1). Furthermore, for fat-tailed infection probabilities ($\alpha < 1$) an excitable phase appears before the transition to endemicity, in which the system exhibits SIR-like dynamics (left panel Fig. 2). Finally, for peaked infection probabilities ($\alpha > 1$) the transition to the endemic state is delayed when approached from below, and additionally becomes discontinuous (right panel Fig. 2).

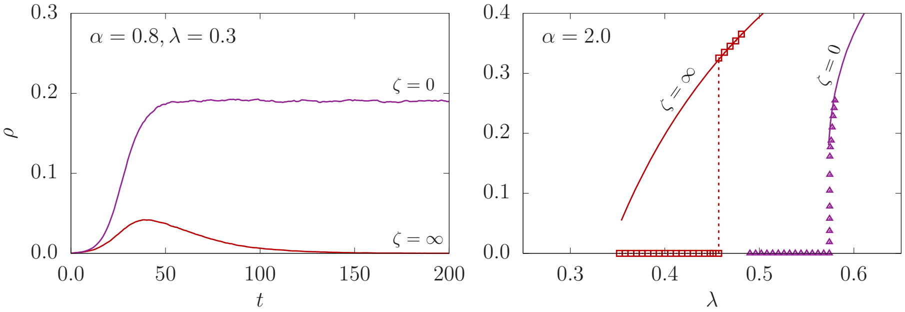


Fig. 2. Left: Averaged prevalence evolution of single infected outbreaks for $\alpha = 0.8$ and $\lambda = 0.3$, corresponding to the endemic phase for short-term memory (purple) and the excitable phase for long-term memory (red). In the endemic phase the outbreak grows monotonically towards its stationary value. In the excitable phase the outbreak infects a large fraction of the population before being eradicated. [Conversely, in the healthy phase the outbreak is eradicated very quickly, infecting only a very small number of individuals.] **Right**: Late-time prevalence when approaching from above (curve) and below (symbols). With short-term memory (purple) both transitions are continuous and occur at the same value of λ. With long-term memory (red) the transition approaching from below is delayed and presents a discontinuous jump.

4 Conclusions

The appearance of this wide array of features, already in unstructured substrates, evidences a crucial role of non-Markovianity in the spread of epidemic outbreaks. The future inclusion of heterogeneous contact networks will shed light on the interplay between memory scales and agent heterogeneities, providing further insight on the relevance of microscopic mechanisms and topological properties in spreading processes.

References

1. Chowell, G., Nishiura, H.: Transmission dynamics and control of ebola virus disease (evd): a review. BMC Medicine 12(1), 196 (Oct 2014), https://doi.org/10.1186/s12916-014-0196-0
2. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature 453(7196), 779–782 (Jun 2008), http://dx.doi.org/10.1038/nature06958
3. Hoffmann, X.R.: Cooperative epidemic spreading. Master's thesis, Universitat de les Illes Balears (Sep 2016), http://ifisc.uib-csic.es/publications/
4. Hoffmann, X.R., Boguñá, M.: In preparation
5. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., Vespignani, A.: Epidemic processes in complex networks. Rev. Mod. Phys. 87, 925–979 (Aug 2015), https://link.aps.org/doi/10.1103/RevModPhys.87.925

Temporal profiles of avalanches on networks

James P. Gleeson[1] and Rick Durrett[2]

[1] MACSI, Department of Mathematics and Statistics, University of Limerick, Ireland
james.gleeson@ul.ie,
WWW home page: http://www.ul.ie/gleesonj
[2] Department of Mathematics, Duke University, Durham, NC, USA

1 Introduction

The dynamics of avalanches or cascades are studied in many disciplines. Examples include the spreading of disease (or information) from human to human [1, 2], avalanches of neuron firings in the brain [3], and the "crackling noise" exhibited by earthquakes and magnetic materials [4]. Of particular interest are cases with dynamics poised at a critical point, where universal scalings of avalanches are observed. The most commonly studied feature of such systems is the distribution of avalanche sizes, which has a power-law scaling at the critical point. The observation of heavy-tailed distributions of avalanche sizes has therefore been used to indicate whether a system is critical. However, power-law distributions can also arise from mechanisms other than criticality [5, 6], so recently attention has focussed more upon the temporal aspects of avalanches, which also exhibit universal characteristics at criticality.

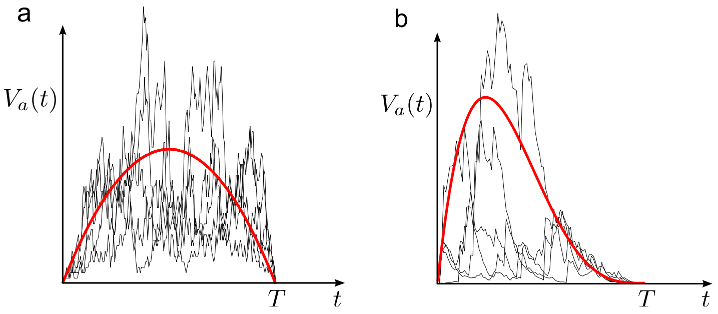


Fig. 1. | **Examples of average avalanche shapes.** In each panel, the black curves show five examples of individual avalanches that all have duration T. The average avalanche shape for the duration T (red curve) is found by averaging the temporal profiles of all such avalanches. Typically, the average avalanche shape is symmetric (e.g., parabolic) as in panel (a), but nonsymmetric avalanche shapes like panel (b) have also been observed (e.g., Fig. S4 of [3]).

The average avalanche shape is determined by averaging the temporal profiles of all avalanches that have a fixed duration T. At criticality, the average avalanche shape is a universal function of the rescaled time t/T, meaning that the average avalanche shapes for different durations can be rescaled to collapse onto a single curve [4]. This feature has recently been used as a sensitive test for criticality in a range of dynamics, from the Barkhausen effect in ferromagnetic materials [7] to neural avalanches [3, 8] and electroencephalography (EEG) recordings from hypoxic neonatal cortex [9]. While average avalanche shapes are typically symmetric (e.g., parabolic) functions of time, nonsymmetric (left-skewed) avalanche shapes have also been observed in experiments. For example, early observations of nonsymmetric avalanche shapes in experiments on Barkhausen noise [4] raised doubts about whether the theoretical model used in [10, 11] was in the correct universality class. Although this discrepancy between theory and experiment was later resolved by a more detailed theory for avalanche propagation [12, 13], several instances of nonsymmetric avalanche shapes (e.g., the neural avalanches in [3]) still lack explanation. Despite some progress in modelling avalanche profiles using random walks [14, 9] and self-organized criticality models [15–18], the factors that cause nonsymmetric average avalanche shapes remain poorly understood.

The characteristics of avalanches that occur on networks depend on both the network connectivity and the node-to-node dynamics of the cascade [19]. Cascading models have been applied, for example, to power-grid blackouts [20], epidemic outbreaks [21], and to the propagation of memes (pieces of digital information) through online social networks [22, 23]. The distribution of avalanche sizes at criticality is known to depend non-trivially on the degree distribution of the underlying network [24], but the time-dependence of cascades has not been studied from this perspective.

In this paper we focus on the temporal profile of cascades, i.e., the average avalanche shape, and how it is affected by the network degree distribution. Using a mathematical derivation of the average avalanche shape for Markovian dynamics (in both critical and noncritical cases) we demonstrate that—as in other universality-breaking examples [25]—networks with heavy-tailed degree distributions can give rise to qualitatively different results from those found on networks with finite-variance degrees. However, the dynamics of the avalanching process are also important: we show that in fact it is the interaction between the dynamics and the network topology that determines whether average avalanche shapes are symmetric or not.

References

1. Pinto, O. A. & Muñoz, M. A. Quasi-neutral theory of epidemic outbreaks. *PloS One*, **6**, e21946 (2011).
2. Borge-Holthoefer, J. *et al.* Cascading behaviour in complex socio-technical networks. *J. Complex Networks* **1**, 3–24 (2013).
3. Friedman, N. *et al.* Universal critical dynamics in high resolution neuronal avalanche data. *Phys. Rev. Lett.* **108**, 208102 (2012).
4. Sethna, J. P., Dahmen, K. A. & Myers, C. R. Crackling noise. *Nature* **410**, 242–250 (2001).
5. Stumpf, M. P. H. & Porter, M. A. Critical truths about power laws. *Science* **335**, 665–666 (2012).
6. Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323–351 (2005).
7. Papanikolaou, S. *et al.* Universality beyond power laws and the average avalanche shape. *Nat. Phys.* **7**, 316–320 (2011).

8. Beggs, J. M. & Timme, N.. Being critical of criticality in the brain. *Front. Physiol.* **3**, 163 (2012).
9. Roberts, J. A. *et al.* Scale-free bursting in human cortex following hypoxia at birth. *J. Neurosci.* **34**, 6557–6572 (2014).
10. Mehta, A. P. *et al.* Universal pulse shape scaling function and exponents: Critical test for avalanche models applied to Barkhausen noise. *Phys. Rev. E* **65**, 046139 (2002).
11. Kuntz, M. C. & Sethna, J. P.. Noise in disordered systems: The power spectrum and dynamic exponents in avalanche models. *Phys. Rev. B* **62**, 11699 (2000).
12. Zapperi, S. *et al.* Signature of effective mass in crackling-noise asymmetry. *Nat. Phys.* **1**, 46–49 (2005).
13. Colaiori, F. Exactly solvable model of avalanches dynamics for Barkhausen crackling noise. *Adv. Phys.* **57**, 287–359 (2008).
14. Baldassarri, A., Colaiori, F. & Castellano, C. Average shape of a fluctuation: Universality in excursions of stochastic processes. *Phys. Rev. Lett.* **90**, 060601 (2003).
15. Laurson, L., Alava, M. J. & Zapperi, S. Power spectra of self-organized critical sandpiles. *J. Stat. Mech.-Theory E* **2005**, L11001 (2005).
16. Rybarsch, M. & Bornholdt, S. Avalanches in self-organized critical neural networks: A minimal model for the neural SOC universality class. *PloS One* **9**, e93090 (2014).
17. Massobrio, P., Pasquale, V. & Martinoia, S. Self-organized criticality in cortical assemblies occurs in concurrent scale-free and small-world networks. *Sci. Rep.* **5**, 10578 (2015).
18. Hesse, J. & Gross, T. Self-organized criticality as a fundamental property of neural systems. *Front. Syst. Neurosci.* **8**, 166 (2014).
19. Larremore, D. B. *et al.* Statistical properties of avalanches in networks. *Phys. Rev. E* **85**, 066131 (2012).
20. Dobson, I. Estimating the propagation and extent of cascading line outages from utility data with a branching process. *IEEE T. Power Syst.* **27**, 2146–2155 (2012).
21. Pastor-Satorras, R. *et al.* Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
22. Cheng, J. *et al.* Can cascades be predicted? In *Proc. 23rd Int. Conf. on World Wide Web*, 925–936 (2014).
23. Goel, S. *et al.* The structural virality of online diffusion. *Manag. Sci.* **62**, 180–196 (2015).
24. Goh, K.-I. *et al.* Sandpile on scale-free networks. *Phys. Rev. Lett.* **91**, 148701 (2003).
25. Radicchi, F. & Castellano, C. Breaking of the site-bond percolation universality in networks. *Nat. Commun.* **6**, 10196 (2015).

Feedback control of the Threshold Model in large-scale networks

Wilbert Samuel Rossi[1], Giacomo Como[2,3], and Fabio Fagnani[2]

[1] Department of Applied Mathematics, University of Twente, The Netherlands,
w.s.rossi@utwente.nl,
[2] Lagrange Department of Mathematical Sciences, Politecnico di Torino, Italy,
giacomo.como@polito.it, fabio.fagnani@polito.it,
[3] Department of Automatic Control, Lund University, Sweden.

Summary. The spread of new behaviors and technologies in social and economic networks are often driven by cascading mechanisms. We consider the Threshold Model of cascades first introduced by Granovetter (1978), where agents choose between two actions following a personal threshold. We allow thresholds to be time-varying and controlled within constraints. We propose a simple feedback mechanism, based on a local mean-field approach, able to reduce the cascade outbreak on large-scale networks.

1 Introduction

Cascading phenomena permeate the dynamics of social and economic networks [1,2,3] [4,5,7,8,9]. One of the most studied models of cascading mechanisms capturing complex neighborhood effects is the Threshold Model (TM) of [3]. Let $\mathscr{N} = (\mathscr{V}, \mathscr{E}, \rho(t), \sigma)$ be a directed network with agent set $\mathscr{V} = \{1, 2, \dots, n\}$ and directed link set $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$. To each agent i are associated a sequence of integer thresholds $\rho_i(t) \in [0, \kappa_i]$, for $t \geq 0$, where κ_i is the out-degree of i, and a binary parameter $\sigma_i \in \{0,1\}$. Agents are endowed with a binary state $Z_i(t) \in \{0,1\}$ initialized by $Z_i(0) = \sigma_i$. The TM says that an agent adopts state-1 if the number of state-1 out-neighbors is at least as large as his current activation threshold; otherwise he adopts state-0:

$$Z_i(t+1) = \begin{cases} 1 \text{ if } & \sum_{j:(i,j)\in E} Z_j(t) \geq \rho_i(t) \\ 0 \text{ otherwise} & \end{cases} \tag{1}$$

Agents apply the local rule (1) synchronously so, given $\mathscr{N}$, the process is determined.

The analysis of the TM is easy for fully mixed populations [3], i.e. for $\mathscr{E} = \mathscr{V} \times \mathscr{V}$. Using the cumulative distribution $F_t(\theta) = n^{-1} \left| \{ i : \rho_i(t) \leq \lfloor \theta n \rfloor \} \right|$, with $0 \leq \theta \leq 1$, and the fraction of state-1 adopters $z(t) = n^{-1} \sum_{i \in \mathscr{V}} Z_i(t)$, the TM dynamics (1) reduces exactly to the scalar recursion:

$$z(t+1) = F_t(z(t)), \quad t \geq 0, \quad \text{with } z(0) = n^{-1} \textstyle\sum_{i \in \mathscr{V}} \sigma_i .$$

2 Local mean-field on regular directed random graphs

We consider the TM on the *ensemble* of all directed networks with size n that have a joint degree/threshold distribution rather than on a specific network $\mathscr{N}$: formally, we consider the directed version of the *configuration model*. In this abstract we restrict to the regular networks, i.e. we assume that the *in*-degree δ_i and the *out*-degree κ_i are k for every node. The fraction of nodes having threshold r at time $t \geq 0$ is

$$p_r(t) = n^{-1}\left|\{i \in \mathcal{V} : \rho_i(t) = r\}\right|, \quad \text{for } 0 \leq r \leq k\,.$$

Although Granovetter's one-dimensional recursion does not hold true in this setting, the fraction of state-1 adopters $z(t)$ in the TM dynamics on most directed networks can be approximated in a quantitatively precise sense by the scalar *local mean-field* recursion

$$x(t+1) = \phi_t(x(t))\,, \qquad x(0) = n^{-1}\textstyle\sum_{i\in\mathcal{V}}\sigma_i$$

where $\phi_t(x)$ is the polynomial with nonnegative coefficients

$$\phi_t(x) := \textstyle\sum_{r=0}^{k} p_r(t)\varphi_{k,r}(x) \quad \text{with} \quad \varphi_{k,r}(x) := \sum_{u=r}^{k}\binom{k}{u}x^u(1-x)^{k-u}\,, \quad 0 \leq r \leq k\,.$$

The recursion was derived in [6] with a concentration result for non-regular networks.

3 Feedback control

Consider a network $\mathcal{N} = (\mathcal{V}, \mathcal{E}, \bar{\rho}, \sigma)$ where the agents have nominal thresholds $\bar{\rho}$ and initial configuration σ such that a cascade of switches to state-1 will occur (the TM can also lead to mixed configurations, or not converge). We are allowed to dynamically modulate the thresholds around $\bar{\rho}$: increasing thresholds makes the agents less sensitive to the number of state-1 neighbors and can prevent a cascade in $\mathcal{N} = (\mathcal{V}, \mathcal{E}, \rho(t), \sigma)$. However, the possible "thresholds increments" are constrained and the optimization of the TM is typically a hard problem. Therefore we propose to identify the nodes to which is most beneficial to increase the thresholds using the local mean-field approach.

We introduce the variables $u_{\bar{r}}^{r}(t) \geq 0$ to describe the fraction of nodes with nominal threshold $\bar{r}$ that at time t have their threshold adjusted to r. These variable need to satisfy $\sum_{r=0}^{k} u_{\bar{r}}^{r}(t) = \bar{p}_{\bar{r}} := n^{-1}\left|\{i \in \mathcal{V} : \bar{\rho}_i = \bar{r}\}\right|$ for $0 \leq \bar{r} \leq k$. The cost to increase the thresholds above their nominal value shall be payed at each time t and is

$$g(\mathbf{u}) = \textstyle\sum_{\bar{r}=0}^{k}\sum_{r=\bar{r}}^{k}\dfrac{|r-\bar{r}|}{k}u_{\bar{r}}^{r}\,.$$

The constraint $g(\mathbf{u}(t)) \leq b$, for every $t \geq 0$, limits the possible threshold increments. Note that at time t the threshold statistic is $p_r(t) = \sum_{\bar{r}=0}^{k} u_{\bar{r}}^{r}(t)$ for $0 \leq r \leq k$.

We propose the following simple feedback scheme to control the TM in the large-scale network. Given the fraction of state-1 adopters $z(t)$ at time t we compute

$$\mathbf{u}^*(t) = \arg\min_{\mathbf{u}} \textstyle\sum_{r=0}^{k}\sum_{\bar{r}=0}^{k} u_{\bar{r}}^{r}\,\varphi_{k,r}(z(t)) \quad \text{subject to} \;\; \sum_{r=0}^{k} u_{\bar{r}}^{r}(t) = \bar{p}_{\bar{r}} \;\text{ and }\; g(\mathbf{u}) \leq b\,.$$

The result $\mathbf{u}^*(t)$ of the Linear Program is then used, up to quantization, to adjust the thresholds of the nodes, from $\bar{\rho}$ to $\rho(t)$. The update of the TM dynamic (1) using $\rho(t)$ will then return the following value $z(t+1)$ of the fraction of state-1 adopters.

Figure 1 contains some simulations over regular directed networks with size $n = 1000$ and in/out-degree $k = 7$. The nominal threshold statistic is

$$\bar{\mathbf{p}} = [0; 0.103; 0.257; 0.128; 0.128; 0.128; 0.128; 0.128]\,,$$

σ is such that $z(0) = 0.5$ and the controlled dynamics has constraint $b = 0.033$.

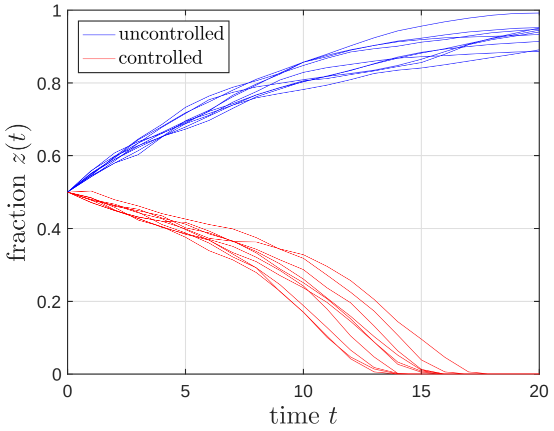


Fig. 1. Simulation example. Ten simulations of the uncontrolled dynamics over ten different randomly chosen networks and ten simulations of the feedback controlled dynamics. In the example, the constraint $b = 0.033$ is sufficient for the feedback control to completely revert the cascade.

4 Conclusion

We propose a simple feedback scheme to control the TM of cascades in large-scale, directed regular networks. The approach uses the fraction of state-1 nodes and the threshold statistics, and can be generalized to non-regular directed graphs with a joint degree/threshold distributions. Further work will focus on investigating the properties of the feedback scheme and of the controlled dynamics.

References

1. Amini, H., Cont, R., Minca, A.: Resilience to contagion in financial networks. Mathematical Finance pp. 1–37 (2013)
2. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, New York, NY, USA (2010)
3. Granovetter, M.: Threshold models of collective behavior. American Journal of Sociology 83(6), 1420–1443 (1978)
4. Jackson, M.O.: Social and Economic Networks. Princeton University Press, Princeton, NJ, USA (2008)
5. Montanari, A., Saberi, A.: The spread of innovations in social networks. Proceedings of the National Academy of Sciences 107(47), 20196–20201 (2010)
6. Rossi, W., Como, G., F.Fagnani: Threshold models of cascades in large-scale networks, https://arxiv.org/abs/1604.05490, submitted, https://arxiv.org/abs/1604.05490
7. Schelling, T.C.: Micromotives and Macrobehavior. W. W. Norton and Company (1978)
8. Vega-Redondo, F.: Complex Social Networks. Econometric Society Monographs, Cambridge University Press (2007)
9. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences 99(9), 5766–5771 (2002), http://www.pnas.org/content/99/9/5766.abstract

Best spreader node in a network

Hale Cetinay, Karel Devriendt, and Piet Van Mieghem

Delft University of Technology, 2600 AA Delft, The Netherlands
{H.Cetinay-iyicil,P.F.A.VanMieghem}@tudelft.nl,
WWW home page: https://www.nas.ewi.tudelft.nl/

1 Introduction

Characterizing a network by a small set of metrics, that are relatively easy to compute and to understand, lies at the heart of network science. Many reviews [1] and books [2] cover graph metrics, real numbers that can be computed from the knowledge of the graph only (e.g. via its adjacency matrix). Each graph metric represents and quantifies a certain property of the graph.

Here, inspired by electrical flows in a resistor network, we propose a promising graph metric that quantifies the nodal spreading capacity in a network, and identifies the best conducting node j in a graph G. This *best spreader node* is found as the minimizer of the diagonal element $Q^{\dagger}_{jj}$ of the pseudo-inverse matrix $Q^{\dagger}$ of the weighted Laplacian matrix of the graph G.

2 Methods

We are interested to find the best spreader node in a network using the pseudo-inverse matrix $Q^{\dagger}$ of the weighted Laplacian $\widetilde{Q}$ of a graph G on N nodes. The major motivation is the appearance of the pseudo-inverse $Q^{\dagger}$ in electrical current flow equations [3]. When a unit current $I_c = 1$ is injected in node j while all others are sinks, the voltage (potential) vector $v = Q^{\dagger}x$ becomes

$$v = Q^{\dagger}\left(e_j - \frac{u}{N}\right) = Q^{\dagger}e_j = \text{col}_j\, Q^{\dagger}$$

where $u = (1,1,\ldots,1)$ is the all-one vector and e_k is the basic vector with the m^{th} component equal to $(e_k)_m = \delta_{mk}$ and δ_{mk} is the Kronecker-delta: $\delta_{mk} = 1$ if $m = k$, otherwise $\delta_{mk} = 0$. Then,

$$v_j = Q^{\dagger}_{jj} \tag{1}$$

is the largest positive potential in $\text{col}_j\, Q^{\dagger}$, as follows physically. When we choose the average potential $v_{av} = \frac{1}{N}\sum_{i=1}^{N} v_i$ equal to zero, we can re-interpret (1) as

$$Q^{\dagger}_{jj} = v_j - v_{av} = \frac{1}{N}\sum_{i=1}^{N}(v_j - v_i)$$

indicating that the best spreader node minimizes the sum of the *potential differences* between its potential v_j and all other nodal potentials.

Node $k^* = \arg_j \left\{ \min_{1 \leq j \leq N} \left(Q^\dagger_{jj} \right) \right\}$ that is electrically best connected to all other nodes, can be regarded as the best diffuser of a flow to the rest of the network, in case a flow (of information or current) is injected in that node. To some extent, node k^* is most influential with respect to the dynamic operations of a network. For instance, in a Markov process, the node k^* in the Markov graph of all states can be regarded as the best, dynamically connected, state, through which the highest probability flux streams towards all other states. In a random walk case, the optimal spreader node k^* possesses the lowest average commute time to all other nodes [4].

3 Results

In Fig. 1, we present the change in the size of the giant component when the network nodes are removed according to five different node-removal strategies. Fig. 1 illustrates that the betweenness, closeness and the diagonal element $Q^\dagger_{jj}$ of the pseudo-inverse matrix $Q^\dagger$ perform similarly in a strategy to disconnect a graph, and question whether a single metric can outperform others in such an NP-complete problem [5]. Sequentially removing the best spreader nodes in the resulting graph is expected a good strategy to

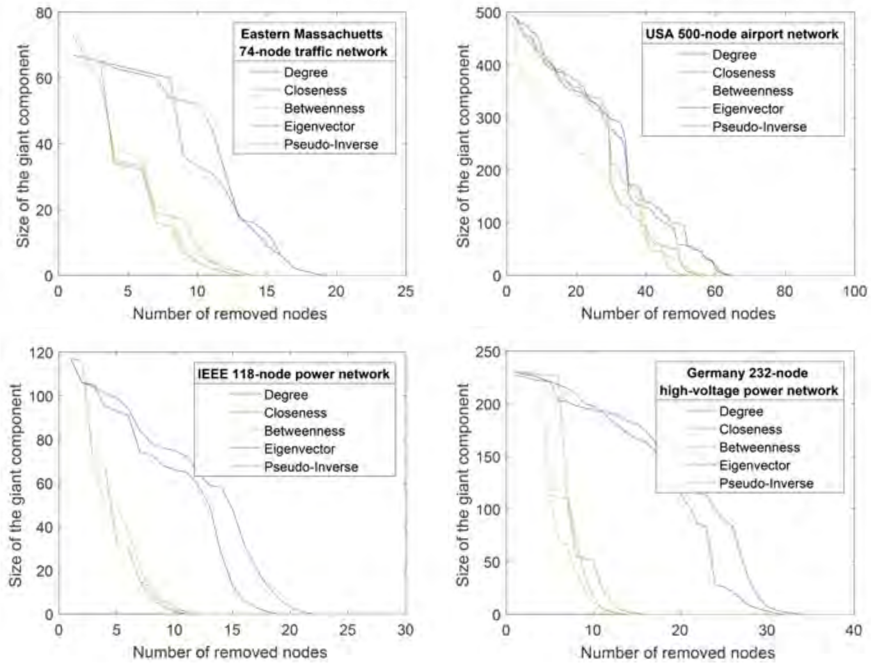


Fig. 1. The size of the giant component in different networks versus the removal of nodes according to five different strategies: the node with the optimal graph metric, computed in each resulting graph, is removed.

fragment the graph. Conversely, protecting the best spreader nodes in a network will result in a robustly designed network.

Summary. Inspired by electrical current flows that satisfy conservation laws, the weighted Laplacian $\widetilde{Q}$ and its pseudo-inverse $Q^{\dagger}$ are argued to be fundamental vehicles to explore properties of graphs as well as dynamic processes in networks. The best spreader, minimum of the diagonal elements of the pseudo-inverse $Q^{\dagger}$, has the lowest energy or potential in the network and the proposed metric opens up applications to various kinds of networks problems.

References

1. J. Martín Hernández and P. Van Mieghem. Classification of graph metrics. Delft University of Technology, Report20111111 (www.nas.ewi.tudelft.nl/people/Piet/TUDelftReports), 2011.
2. P. Van Mieghem. *Performance Analysis of Complex Networks and Systems*. Cambridge University Press, Cambridge, U.K., 2014.
3. H. Cetinay, F. A. Kuipers, and P. Van Mieghem. A topological investigation of power flow. *IEEE Systems Journal*, to appear 2016.
4. P. Van Mieghem, K. Devriendt and H. Cetinay, Pseudo-inverse of the Laplacian and best spreader node in a network. *Physical Review E*, vol. 96, No. 3, p 032311.
5. P. Van Mieghem, D. Stevanović, F. A. Kuipers, C. Li, R. van de Bovenkamp, D. Liu, and H. Wang. Decreasing the spectral radius of a graph by link removals. *Physical Review E*, 84(1):016101, July 2011.

Infection Spreading in Temporal Networks With Memory

Oliver E. Williams[1], Fabrizio Lillo[2], and Vito Latora[1]

[1] School of Mathematical Sciences, Queen Mary University of London, London, E1 4NS, United Kingdom
[2] Department of Mathematics, University of Bologna, Piazza di Porta San Donato 5, 40126, Bologna, Italy

1 Introduction

The importance of memory in the processes that underlie many types of real-world systems is obvious. One would not expect individuals at a party to randomly choose who to talk to at any point in time, but rather have a group of friends that they keep coming back to, just as we would not expect the destinations of travellers at an airport to be chosen purely based on the location of the airport.[5] We expect these systems to have memory; their past is important in determining their future. This non-Markovian nature can be introduced into models for such systems in a number of ways. When it is introduced we expect there to be some influence on any processes running on the network.[3][7] Studies of flows in various real-world networks have shown that simple Markov models do not capture many important aspects of the systems in question.[6][5] When only a single step of memory is taken into account (i.e. the spreading can be described by a second order Markov process) one can observe either the slow down, or speed up, of information spreading across a network.[1][7] Work on the spreading of infection in a time varying network with non-exponential contact times has also shown that memory can significantly effect the epidemic threshold of a SIS model.[8] What we aim to study is the effect that changing the extent of this memory has on the spread of an epidemic.

2 The DAR(p) Network

We propose a method to construct a time-varying network in discrete time: each link from node to node will be generated by its own independent stochastic process. More formally we say that given a set of nodes V we assign to each possible pair $v_i, v_j \in V$ a discrete time stochastic process X_t^{ij} such that $X_t^{ij} \in \{0,1\} \forall t$. In this way the adjacency matrix for the network is defined entry by entry. For our purposes, we take links to be undirected, and we take each X_t^{ij} to be independent and identically distributed and as such we can talk more generally about a process X_t without worrying about which link we are referring to. The particular process X_t is chosen as a special case of the discrete auto-regressive process of fixed order p, hereafter referred to as $DAR(p)$.[2] The principle here can be explained as follows: first we randomly choose whether to draw a link randomly or not, if not then uniformly pick a state in the last p steps of the time series, otherwise the link exists with probability y. In terms of random variables this can be written as:

$$X_t = V_t X_{(t-Z_t)} + (1 - V_t)Y_t. \tag{1}$$

where $V_t \sim Bernoulli(q)$, $Y_t \sim Bernoulli(y)$ and Z_t is some random variable which picks integers in the range $(1, ..., p)$. In principle this Z_t could take any form, but for the sake of simplicity we here take $Z_t \sim Uniform(1, p)$. These modelling choices ensure that $X_t \in \{0, 1\} \forall t$, and so the adjacency matrix generated at each time t will be unweighted. This model allows precise control over the memory in the network.

3 Infection Spreading

We consider the simplest possible mechanism for propagating a disease (or some "message") through the temporal networks defined above: the SI model (a special case of the SIS model, but with the recovery rate set to zero).[4] Here the model is interpreted as follows: if we define I_t to be the set of infected nodes at a time t, and take some initial subset $I_0 \subset V$ of nodes in the infected state, then at each time t for all infected nodes $v_i \in I_t$ each neighbouring node $v_j \in \partial v_i(t)$ (where $\partial x(t) := \left\{ y \in V : A_{xy}(t) = 1 \right\}$) becomes infected with probability λ. Since there is no recovery, this change in state is permanent. For the purposes of simulation we always start with $|I_0| = 1$.

4 Results

An analytical expression for the average time taken for an infection to pass along a link in this network ($\langle \tau \rangle_p$) has been found by casting the process as a p_{th} order Markov chain. This result is then used to show that, for some values of λ, q and y there exists a non-trivial memory length $p = p_{crit}$ such that $\max_p \left\{ \langle \tau \rangle_p \right\} = \langle \tau \rangle_{p_{crit}}$. This inflection point is then observed in the average time taken to infect a full network, or the average time taken for an infection to pass between any two nodes in a full network (as generated by the $DAR(p)$ process).

References

1. Gueuning, M., Delvenne, J.-C., Lambiotte, R.: Imperfect spreading on temporal networks. Eur. Phys. J. B 88(11), 282 (2015), https://doi.org/10.1140/epjb/e2015-60596-0
2. Jacobs, P.A., Lewis, P.A.: Discrete time series generated by mixtures. iii. autoregressive processes (dar (p)). Tech. rep., NAVAL POSTGRADUATE SCHOOL MONTEREY CALIF (1978)
3. Lambiotte, R., Salnikov, V., Rosvall, M.: Effect of memory on the dynamics of random walks on networks. Journal of Complex Networks 3(2), 177–188 (2015), + http://dx.doi.org/10.1093/comnet/cnu017
4. Prakash, B.A., Tong, H., Valler, N., Faloutsos, M., Faloutsos, C.: Virus Propagation on Time-Varying Networks: Theory and Immunization Algorithms, pp. 99–114. Springer Berlin Heidelberg, Berlin, Heidelberg (2010), https://doi.org/10.1007/978-3-642-15939-8_7

5. Rosvall, M., Esquivel, A.V., Lancichinetti, A., West, J.D., Lambiotte, R.: Memory in network flows and its effects on spreading dynamics and community detection. Nat. Commun. 5, 4630 (08 2014), http://dx.doi.org/10.1038/ncomms5630
6. Salnikov, V., Schaub, M.T., Lambiotte, R.: Using higher-order markov models to reveal flow-based communities in networks. Sci. Rep. 6, 23194 (03 2016), http://dx.doi.org/10.1038/srep23194
7. Scholtes, I., Wider, N., Pfitzner, R., Garas, A., Tessone, C.J., Schweitzer, F.: Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. Nat. Commun. 5, 5024 (09 2014), http://dx.doi.org/10.1038/ncomms6024
8. Van Mieghem, P., van de Bovenkamp, R.: Non-markovian infection spread dramatically alters the susceptible-infected-susceptible epidemic threshold in networks. Phys. Rev. Lett. 110, 108701 (Mar 2013), https://link.aps.org/doi/10.1103/PhysRevLett.110.108701

Ranking of Nodal Infection Probability in Susceptible-Infected-Susceptible Epidemic

Bo Qu[1], Cong Li[2], Piet Van Mieghem[1], and Huijuan Wang[1]

[1] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, 2624CZ, The Netherlands,
[2] School of Information Science and Engineering, Fudan University, Shanghai 200433, China,

The Susceptible-infected-susceptible (SIS) epidemic process has been widely studied as a model of virus spread on a network [2,9,4,1,8,11,7,5]. In the SIS model, a node is either infected or susceptible at any time t. Each infected node infects each of its susceptible neighbors with an infection rate β. Each infected node recovers with a recovery rate δ. Both infection and recovery processes are independent Poisson processes and the ratio $\tau = \beta/\delta$ is the effective infection rate. There is an epidemic threshold τ_c and a nonzero fraction of nodes is infected in the metastable state when the effective infection rate is above the threshold $\tau > \tau_c$. The infection probability $v_{k\infty}(\tau)$ of a node k in the metastable state at a given effective infection rate τ indicates the vulnerability of node k to the virus, and the average fraction $y_\infty(\tau)$ of infected nodes reflects the global vulnerability of the network.

Researchers have mainly concentrated on the average fraction y_∞ of infected nodes in the metastable state to estimate the vulnerability of a network against a certain epidemic or virus. Great effort has been devoted to understand how the network topology influences the vulnerability and the epidemic threshold[5,9]. The nodal vulnerability or equivalently the infection probability of each node in the metastable state, however, has been seldom studied, except for special cases, i.e. when the effective infection rate is just above the epidemic threshold [10]. In this case, it is found that, the metastable-state infection probability vector $V_\infty = [v_{1\infty} v_{2\infty} \cdot\ \cdot\ \cdot v_{N\infty}]^T$), obtained by the N-Intertwined Mean-Field Approximation (NIMFA) of SIS model is proportional to the principal eigenvector x_1 of the adjacency matrix A.

In this work, we aim to explore the nodal infection probability in a systematic way, in different network topologies and when the effective infection rate τ varies. As a starting point, we investigate the ranking of nodal infection probabilities, which crucially informs a network operator which nodes are more vulnerable or require protection. Interestingly, we find that the ranking of the nodal infection probability changes as the effective infection rate τ varies. The observation points out that we cannot find a topological feature of a node to represent the vulnerability of a node to an SIS epidemic, because the rankings in vulnerability of nodes in a network may be different when the effective infection rate τ varies, whereas a topological feature of a node remains the same. Our observation explains the finding of Hebert-Dufresne et al. [3] that different nodal features (such as degree, betweenness, etc.) should be used to select the nodes to immunize in different scenarios (based on different infection rates, link densities, etc.), i.e. different nodes should be immunized at different infection rates. In this paper, we explore two questions: 1. in which network topology the ranking of nodal infection probabilities changes more significantly when the effective infection rate τ varies and

2. in which effective infection rate range, the increment of the effective infection rate leads to a more significant change in the ranking for a given network topology? Via both theoretical and numerical approaches, we unveil that the ranking of nodal vulnerability changes more dramatically when τ is smaller or in Barabási-Albert than Erdős-Rényi random graphs.

For more information, we refer to [6]

References

1. Boccara, N., Cheong, K.: Critical behaviour of a probabilistic automata network SIS model for the spread of an infectious disease in a population of moving individuals. J Phys A-Math Gen 26(15), 3707 (1993)
2. Daley, D.J., Gani, J., Gani, J.M.: Epidemic modelling: an introduction, vol. 15. Cambridge University Press (2001)
3. Hébert-Dufresne, L., Allard, A., Young, J.G., Dubé, L.J.: Global efficiency of local immunization on complex networks. Scientific Reports 3 (2013)
4. Li, C., van de Bovenkamp, R., Van Mieghem, P.: Susceptible-infected-susceptible model: A comparison of n-intertwined and heterogeneous mean-field approximations. Phys. Rev. E 86(2), 026116 (2012)
5. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., Vespignani, A.: Epidemic processes in complex networks. Rev. Mod. Phys. 87, 925–979 (Aug 2015), https://link.aps.org/doi/10.1103/RevModPhys.87.925
6. Qu, B., Li, C., Van Mieghem, P., Wang, H.: Ranking of nodal infection probability in susceptible-infected-susceptible epidemic. Scientific Reports 7(9233) (2017)
7. Qu, B., Wang, H.: Sis epidemic spreading with heterogeneous infection rates. IEEE Transactions on Network Science and Engineering 4(3), 177 – 186 (2017)
8. Shi, H., Duan, Z., Chen, G.: An SIS model with infective medium on complex networks. Physica A 387(8), 2133–2144 (2008)
9. Van Mieghem, P.: The N-intertwined SIS epidemic network model. Computing 93(2-4), 147–169 (2011)
10. Van Mieghem, P.: Performance analysis of communications networks and systems. Cambridge University Press (2014)
11. Wang, H., Li, Q., DAgostino, G., Havlin, S., Stanley, H.E., Van Mieghem, P.: Effect of the interconnected network structure on the epidemic threshold. Physical Review E 88(2), 022801 (2013)

Speeding up non-Markovian SI spreading with a few extra edges

Alexey Medvedev[1,2] and Gábor Pete[3,4]

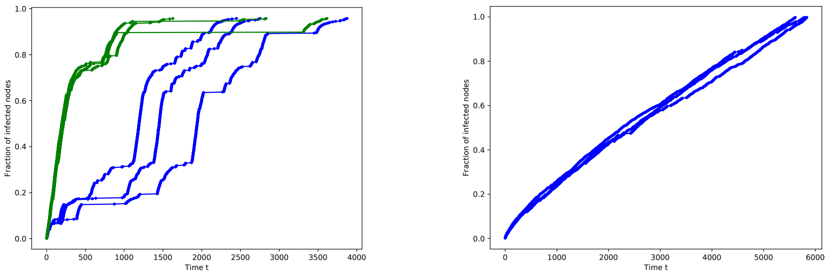
[1] NaXys, Université de Namur, Namur, B-5000, Belgium
[2] ICTEAM, Université Catholique de Louvain, Louvain-la-Neuve, B-1348, Belgium
[3] Alfréd Rényi Institute of Mathematics, 1053, Budapest, Hungary
[4] Budapest University of Technology and Economics, 1111, Budapest, Hungary
an_medvedev@yahoo.com,
WWW home page: http://alexeymedvedev.com

1 Introduction and motivation

Spreading is one of the most important dynamic processes on complex networks as it is the basis of a broad range of phenomena from epidemic contagion to diffusion of innovations. One of the original, and still primary, reasons for studying networks is to understand the mechanisms by which diseases, information, computer viruses, rumors, innovations spread over them. We consider the two-state *susceptible-infected (SI) model* on the rooted connected simple graph $G = (V, E, s)$, where vertices can be either in susceptible (S) or in infected (I) state. Infection is then transmitted along the edges with the transition rule S → I, meaning that once an infection is obtained, the vertex stays infected forever. Initially, only the root $s \in V$ is infected. The edges $e \in E$ have i.i.d. random positive weights $\xi(e)$, with common distribution ξ, representing the passage time of the infection. Then, we can measure spreading via stochastic process $(T_k)_{k=1}^{|V|}$, where T_k denotes the time to infect k vertices in the graph.

Social interactions often follow bursty patterns, which are usually modelled with non-Markovian heavy-tailed edge weights, therefore we let ξ follow *power law distribution* $\mathsf{pow}(\alpha)$: $\mathbb{P}(\xi > t) = 1 \wedge (t/t_0)^{-\alpha}$, where $t_0 > 0$ and $\alpha > 0$. Most sparse (e.g., bounded average degree) random graph models produce graphs that are *locally tree-like*: a large neighbourhood of a random root is a tree with high probability, maybe with some extra edges decorating the tree. Typically, *supercritical* random graph models have been studied, where the unique giant connected component looks locally like a fast-growing supercritical Galton-Watson tree, maybe with decorations. This local fast-growing tree structure has been used in several works very successfully to understand the behavior of SI on the entire graph [2, 1].

However, it was observed in [6] that the SI spreading with heavy-tailed ξ on *finite trees*, when started from a typical site, may be very slow. Real-life examples with such ξ include, for example, cascades of retweets in Twitter [4, 3]. A striking feature of slow spreading when ξ has infinite mean was noticed in [5] via computer simulations, which is most apparent when curve $k \mapsto \langle T_k \rangle$ is considered. Namely, running the simulation of the process $T = (T_k)_{k=1}^{n}$ on a tree for M times, and defininig $\langle T_k \rangle = \frac{1}{M}\sum_{i=1}^{M} T_k^{(i)}$, the *spreading curve* $(\langle T_k \rangle, k/n)$ exhibits "uncontrolled large plateaux", which do not



(a)

(b)

Fig. 1: Spreading curves of the SI model simulation with power law weights ξ with $\alpha = 0.8$. (a) For the three dark blue curves, the underlying graph is a tree with 472 vertices; for the three lighter green curves, one fixed edge is added between the root and a random vertex; b) the underlying graph is a cycle C_n with 472 vertices.

decrease and do not converge as we increase the number of runs (see the dark blue curves on Figure 1 (a)). Plateaux of similar type were also empirically found in [3]. On the other hand, adding just one extra edge to the tree (*which does not change the local statistics of the graph!*) makes the spreading curve quite smooth; see the lighter green curves on Figure 1 (a), and also Figure 1 (b) that shows SI spreading on the cycle, with power law exponent $\alpha \in (1/2, 1)$.

Summarizing the above assumptions, in this paper we consider the SI spreading model with i.i.d. power law transmission times and we study the role of cycles in speeding up of the process when $\alpha \in (1/2, 1)$ and thus eliminating large plateaux on (or 'smooth') the spreading curve up to a certain level. We rigorously identify when the first possible temporal bottleneck appears and its relation to the graph structure. We consider two natural models of random trees where we study the striking effect how adding a few edges typically smooths spreading curve and conclude by studying the case on the largest cluster in a *near-critical Erdős-Rényi graph* $G\big(n, \frac{1}{n} + \frac{\lambda}{n^{4/3}}\big)$.

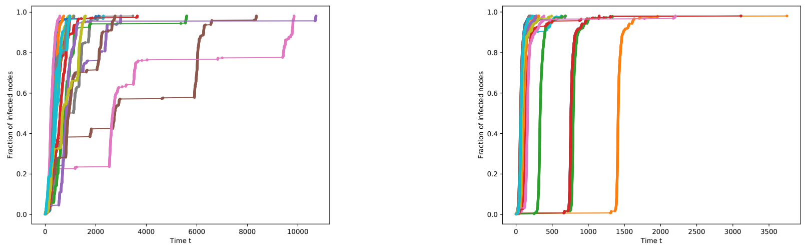
2 Results

Here we give a novel presentation of the results. We show that for each finite connected graph G on n vertices there exists a specific threshold $\kappa(G,s)$, where s is an initially infected vertex, such that the average time to infect $k \leq \kappa(G,s)$ vertices is finite, and is even bounded by a polynomial function in k, and when $k > \kappa(G,s)$ the average time is infinite. The number $\kappa(G,s)$ identifies the position of the first temporal bottleneck for the process, and has a simple combinatorial description, which is presented in the following Lemma.

Lemma 1. *Let G be a finite rooted graph with root s and let T be the SI spreading process on G with weights having absolutely continuous distribution ξ, such that $\mathbb{E}(\xi) = \infty$. Then,*

$$\kappa(G,s) = \min_{e \in E(G)} |\mathscr{C}(s, G \backslash e)|,$$

where $|\mathscr{C}(s, G \backslash e)|$ is the size of the connected component of vertex s in the graph G without edge e.



(a)

(b)

Fig. 2: Simulation of SI spreading with power law inter-event times with $\alpha = 0.8$ on the largest component of a near-critical Erdős-Rényi graph with n vertices. The edge weights are fixed, 20 runs are shown with random starting vertices. (a) $\lambda = 0$, $n = 6000$ and the component has 541 vertices, no surplus edges. (b) $\lambda = 5$, $n = 1000$ the component has 515 vertices, 27 surplus edges.

Next we study the value of $\kappa(G,s)$ in some near-critical random graph models, and show that by adding one or a few edges, it typically increases from a bounded value to a positive proportion of all the vertices. In our first example, we add an extra edge to a *critical Galton-Watson tree* conditioned to have depth at least N, between the starting vertex s and a uniform random vertex. In our second example, we add one random edge to the *uniform spanning tree* of the complete graph. In order to significantly raise $\kappa(G,s)$, it is important that the extra edge is incident to s, or in other words, that s lies in the cycle that we have created. Otherwise, similarly to adding a link, we may start the infection from two random points to achieve the same effect.

Finally, our third example is the "physically" most relevant one: the *near-critical Erdős-Rényi graph* $G\big(n, \frac{1}{n} + \frac{\lambda}{n^{4/3}}\big)$, $\lambda \in \mathbb{R}$. For $\lambda \ll 0$ it is very likely to be a large critical tree, while for $\lambda \gg 0$ it is very likely to be a critical tree with several extra edges. Here, if the infection is started from a uniformly random vertex σ, the $\kappa(G,\sigma)$ will be a bounded random variable, thus plateaux will emerge immediately. However, after infecting just a tiny portion of the graph, the infection whp reaches the point s, such that $\kappa(G,s)$ is large and later will spread up to almost the whole graph in finite expected time (see Figure 2).

References

1. Baroni, E., Hofstad, R.v.d., Komjáthy, J.: Nonuniversality of weighted random graphs with infinite variance degree. Journal of Applied Probability **54**(1) (2017) 146–164
2. Bhamidi, S., van der Hofstad, R.v.d., Komjáthy, J.: The front of the epidemic spread and first passage percolation. Journal of Applied Probability **51A** (12 2014) 101–121
3. Bhowmick, A.K., Gueuning, M., Delvenne, J.C., Lambiotte, R., Mitra, B.: Temporal pattern of (re)tweets reveal cascade migration. In: Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM 2017, Sydney, Australia. (2017) **to appear**.
4. Gandica, Y., Carvalho, J., Sampaio dos Aidos, F., Lambiotte, R., Carletti, T.: Stationarity of the inter-event power-law distributions. PLOS ONE **12**(3) (03 2017) 1–10
5. Horvth, D.X., Kertsz, J.: Spreading dynamics on networks: the role of burstiness, topology and non-stationarity. New Journal of Physics **16**(7) (2014) 073037
6. Iribarren, J.L., Moro, E.: Impact of human activity patterns on the dynamics of information diffusion. Phys. Rev. Lett. **103** (Jul 2009) 038702

Crawling in crowed conditions. Application to network reconstruction.

Timoteo Carletti[1], Malbor Asllani[1], Francesca Di Patti[2], Duccio Fanelli[2] and Francesco Piazza[3]

[1] naXys, Namur Institute for Complex Systems, University of Namur, Belgium,
timoteo.carletti@unamur.be,
WWW home page: https://directory.unamur.be/staff/tcarlett
[2] Dipartimento di Fisica e Astronomia, Università degli Studi di Firenze and INFN, Sezione di Firenze, Italy
[3] Centre de Biophysique Moléculaire, CNRS-UPR 4301, Université d'Orléans, France

1 Introduction

Diffusion is one of most studied problem in nature. It investigates the transport phenomena due the spontaneous spreading of mass in a defined domain. The first empirical description of this process was formulated as the Fick's law of diffusion [1]. Successively, thanks first to Einstein [2] and then to Smoluchowski [3], the diffusion process was analytically modelled on the basis of previous brownian motion observations [4].

From then on, diffusion-like processes were applied to many areas, from physics to biology [5–7]. Moreover in the last years, attention has been focused on the dynamics of particles on specific discrete mediums, namely the complex networks. This interest rises from the similarity between the topological properties of these structures and those characterizing real life phenomena, such as cell compartments or traffic flow [8]. Networks have nowadays a central role in the modelling of physical problems, and thus the diffusion on such environments has become an emergent field of investigation [9–13]. Most of the studies done in this area focussed on the detection of community structures [10, 11, 14], or on the prediction of stationary patterns [15], and finally on the problem of diffusive epidemic process in a crowded environment [16].

2 Model

In all the previously presented models, it has been hypothesized the existence of a single walker or independent ones if many were active. In this work, we make one step forward by studying the simultaneous diffusion of several random walkers on a complex symmetric network in a crowded regime. More precisely, each individual crawler sitting on a node obeys the standard rules of random walk, jumps to distance 1 nodes are equiprobable, however each node can account for only a finite number of walkers (N), namely it possesses a finite carrying capacity. This implies that the transition probability from one node to another takes also into consideration the quantity of free available volume in the destination nodes.

Starting from this microscopic formulation, we derive a diffusion equation characterized by a transport operator that differs from the standard random-walk Laplacian matrix, notably for the presence of nonlinear terms involving products of the density of walkers:

$$\frac{\partial}{\partial t}\rho_i = \sum_{j=1}^{\Omega} \Delta_{ij}\left[\rho_j\,(1-\rho_i) - \frac{k_j}{k_i}\rho_i\,(1-\rho_j)\right] \tag{1}$$

where A_{ij} is the $\Omega \times \Omega$ symmetric adjacency matrix encoding the connections in the network, $k_i = \sum_j A_{ij}$ the degree of the i–th node, $\Delta_{ij} = A_{ij}/k_j - \delta_{ij}$ the random-walk normalized Laplacian matrix and the continuous variable $\rho_i(t)$ represents the concentration of particles at node i and it is related to the discrete variable n_i (number of walkers at node i) through $\rho_i = \lim_{N\to\infty}\langle n_i\rangle/N$.

The different structure of the diffusion is directly reflected on the stationary solution: the asymptotic concentration of crawlers in each node is no longer proportional to the degree of the node itself but it is given by a nonlinear function of the degree. We are able to derive an analytical formula for such solution valid for any crowding conditions and which returns the classical random walk distribution in the case of diluted systems, namely once the number of crawlers is very small with respect to the available volume:

$$\rho_i^{\infty} = \frac{ak_i}{1+ak_i} \quad \forall i = 1,\ldots,\Omega\,, \tag{2}$$

where a is a parameter to be determined to satisfy the mass conservation constraint, $\sum_i \rho_i(t) = \sum_i \rho_i(0) = M$, which straightforwardly follows from Eq. (1).

3 Results

We conclude by presenting a main application of the previous theory devoted to the reconstruction of the unknown network topology upon which the crawlers move, remarkably enough a *single node measurement* is sufficient to achieve the goal. We are indeed able to reconstruct the degree distribution $p(k)$ using the information on $\rho_i(t)$ (with t large enough) observed on a single node and repeating s independent experiments involving different numbers of crawlers (whatever the crowding conditions).

More precisely selecting a node as starting point for all the walkers, say node $i = 1$, we can use Eq. (2) and get $a(M) = \rho_1^{\infty}/(1-\rho_1^{\infty}) \times 1/k_1$, where we emphasised the dependence on the *system mass* M, i.e. the total number of walker. Let us observe that $a(M)$ depends only on local measurable quantities: the node degree k_1 and the stationary distribution of walkers on the node, ρ_1^{∞}, that one can safely assume to be know if one waits long enough observing the number of walkers contained in node 1.

Introducing the number of nodes with degree k, $n(k)$, we can obtain from Eq. (2)

$$M = \sum_k n(k)\frac{a(M)k}{1+a(M)k}\,, \tag{3}$$

performing several *experiments*, namely random walks with a different number of walkers M_i, $i = 1, \dots, s$, one can rewrite the previous relation in the following form:

$$\begin{pmatrix} M_1 \\ \vdots \\ M_s \end{pmatrix} = F \begin{pmatrix} n(1) \\ \vdots \\ n(k_{max}) \end{pmatrix} , \tag{4}$$

where we introduced the matrix $F_{ij} = \frac{a_i j}{1 + a_i j}$ and we wrote for short $a_i = a(M_i)$, that we recall is a known quantity.

Solving this linear system for the unknown $n(1), \dots, n(k_{max})$ we can reconstruct the degree distribution of the network. We successfully tested our method in both synthetic networks (see Fig. 1 for the case of an Erdős-Rény network and a Scale Free one) but also in realistic ones (*C. Elegans* neural network and the *karate club* network, data not shown).

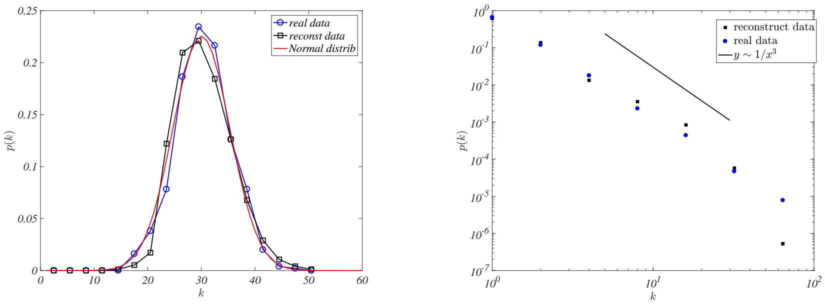


Fig. 1. Network reconstruction. Left panel: the degree distribution $p(k)$ for an Erdős-Rény random network made by $\Omega = 500$ nodes and probability to have a link between two nodes $p = 0.06$. Right panel: The degree distribution $p(k)$ for a Scale Free network made by $\Omega = 2000$ nodes and $\gamma = 3$ built using the Barabási-Albert algorithm. In both panels the blue circles denote the real probability distribution (i.e. computed from the knowledge of the network) while the black squares represent the reconstructed $p(k)$, the red line (left panel) is the asymptotic binomial distribution with parameters Ω and p, the black line (right panel) is the theoretical distribution $y \sim 1/x^3$.

References

1. Crank J., The Mathematics of Diffusion, 2rd ed. (OUP, Oxford) (1975)
2. Einstein A., Annalen der Physik, Vol. 17 (1905), p. 549.
3. von Smoluchowski M., Annalen der Physik, Vol. 21 (1906), p. 756.
4. Brown R., Phil. Mag., Vol. 4 (1828), p. 161.
5. Bouchaud R. and Georges A., Phys. Rep., Vol. 195 (1990), p. 127.
6. Zaslavsky G.M., Phys. Rep., Vol. 371 (2002), p. 461.

7. Metzeler R. and Klafter J., Phys. Rep., Vol. 339 (2000), p. 1.
8. Barrat A., Barthélemy M. and Vespignani A., Dynamical Processes on Complex Networks, 1st ed. (Cambridge University Press, Cambridge) (2008)
9. Almaas E., Kulkarni R.V. and Stroud D., Phys. Rev. E, Vol. 68 (2003), p. 056105.
10. Simonsen I., Eriksen K.A., Maslov S. and Sneppen K., Physica A, Vol. 336 (2004), p. 163.
11. Simonsen I., Physica A, Vol. 357 (2005), p. 317.
12. Burioni R., Chibbaro S., Vergni D. and Vulpiani A., Phys. Rev. E, Vol. 86 (2012), p. 055101(R).
13. Masuda N., Porter M.A. and Lambiotte R., Physics Reports, August (2017)
14. Lambiotte, R., Delvenne, J. C. and Barahona, M., IEEE Transactions on Network Science and Engineering, Vol. 1 (2) (2014) , p. 76.
15. Angstmann C.N., Donnelly I.C. and and Henry B.I., Phys. Rev. E, Vol. 87 (2013), p. 032804.
16. Kwon S. and Kim Y., Phys. Rev. E, Vol. 84 (2011), p. 041103.

Hierarchy measurement for modeling network dynamics under directed attacks

Rubinson M.[1], Levit-Binnun N.[2], Peled A.[3,4], Naim-Feil J.[1,2], Freche D.[1,2], Moses E.[1]

[1] Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel
mica.rubinson@weizmann.ac.il
[2] Sagol Center for Brain and Mind, Baruch Ivcher School of Psychology, IDC, Herzliya, Israel
[3] Sha'ar Menashe Mental Health Center, Sha'ar Menashe, Israel
[4] Ruth and Bruce Rappaport Faculty of Medicine, Technion, Haifa, Israel

1 Introduction

Structural graph properties are largely responsible for the way information flows through the network [1]. Indeed, a network's description is not complete without considering its dynamic responses, such as the effects of rewiring links or the insertion and removal of nodes over time. In this research we focus on two types of networks' dynamics – attack and activation, and investigate how they relate to each other. In the attack process, removal of nodes induces a cascade of abrupt dynamic failures, which eventually cause the network to collapse [2]. Activation is an aspect of percolation dynamics, which measures the efficacy of spreading a signal in the network. In a simple spreading activation model, S_i is the activation of node V_i which accumulates at each time step, and the sum of all activations will grow exponentially $\sum_i S_i \sim \exp(\mu_s t)$, with an activation exponent μ_s that is dependent on the graph's topology [3].

Our approach describes the dynamic process from the graph's static configuration, which can potentially be used for real-time intervention. We complement global aspects of network behavior by focusing on an important local aspect: the removal order of the nodes, which is dependent on the system's dynamics. While local characteristics such as the nodal degree and clustering coefficient can identify dominant members in the network, they are often ignored in the study of network dynamics and averaged into global parameters.

2 The *h* function

In the process of a degree-based attack, the time step $t=(1,\ldots,N)$ of removal from the network imposes an order on the nodes V_i, different than that imposed by the index i. After scaling by N to get $\tau \equiv t/N$, the correspondence between i and τ is obtained by defining τ_i to be the time step at which node V_i is removed. Its degree at that time, defined as M_i, is then the maximal degree. While k_i is defined by the initial graph, M_i depends on the details of the removal process. In an effort to predict the removal order, one can consider the original degrees of the nodes $\{k_\text{i}|\text{i}=1,\ldots,N\}$ as a first order approximation. However, this turns out to be a rather rough and inaccurate approximation, since the attack process terminates connections and therefore changes the degrees of the remaining nodes. We present a second order approximation that is dependent both on the degree of the node itself and on the degrees of its nearest neighbors (NN):

$$h_i = N_{i-} - N_{i+} \tag{1}$$

Where N_{i-} is the number of NN that are less connected than V_i and N_{i+} is the number of NN that are more connected than V_i.

3 Datasets and connectivity graphs

We tested our approach and present our results using a variety of both simulated and experimentally measured graphs:

- **Erdös-Renyi [4] (ER) Random graphs and Barabasi-Albert (BA) [5] Scale-free graphs:** For both types of topologies, 100 graphs for each size N=(100, 200,…, 1000) were generated. The average degree of all graphs was fixed to $<k>=10$. The size range was chosen based on co criteria of convergence.
- **Partially randomized Scale-free graphs:** 100 BA scale-free graphs of sizes N=(100, 200,…, 1000) were randomized such that a "randomization level" p [%] of the edges were rewired randomly. p ranges between 5% and 100%. The size range was chosen based on criteria of convergence.
- **EEG graphs:** 64-channel EEG recordings of 20 healthy subjects that included 5 seconds of eyes-open resting state were acquired. Data was pre-cleaned using standard EEG analysis protocols and connectivity matrices were computed according to a "Gradient Montage" [6]. A threshold was applied such that all graphs are binary and have a fixed mean degree $<k>=20$ [7].

4 Results and discussion

We find that the maximal degree in the network $M(\tau)$ decays exponentially with rescaled time τ: $M(\tau)\sim exp(-\mu_k\,\tau)$, where μ_k is a topology-dependent attack exponent. To compare whether k_i or h_i better predicts the removal order τ_i, we use the Spearman ranked correlation coefficient (SRCC) between τ_i and each predictor for varying percentages of removed nodes. Denoting r_h and r_k as the SRCC for h vs. τ and k vs. τ, respectively, we find that the h-based prediction is always superior to the k-based prediction (Fig. 1). This goes beyond well-defined graph topologies, and applies to real human brain EEG networks that are neither scale-free nor ER-random.

We go on to investigate the link between exponential behaviors in the attack and activation dynamics, utilizing partially randomized scale-free graphs. For given N and $<k>$ and varying p, the spreading activation exponent μ_s and the attack exponent μ_k were calculated for each graph (averaged over 100 randomizations) and shown to decrease monotonically with the level of randomness. An increase in μ_k values corresponds to faster collapse of the graph, while higher μ_s values correspond to more efficient spreading of activation in the graph. These exponents thus shed light on the degree of hierarchy in the graph topology. Furthermore, μ_k and μ_s are found to strongly correlate with each other (Fig. 2) and are indicators of the graph's level of randomness. These results emphasize the connection between the failure and activation of networks, and use dynamics to distinguish between different topologies.

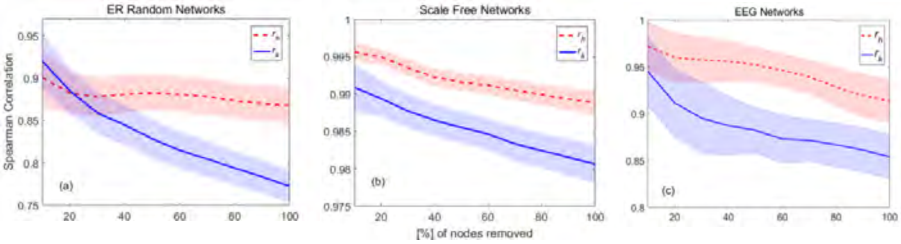
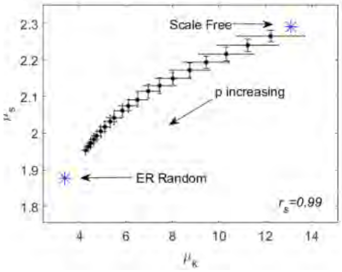


Fig. 2. Spearman ranked correlation between the removal order τ and the initial degree k (blue line) or the hierarchy function h (red dashed line), plotted with respect to the percentage of nodes removed: (a) ER random networks (N=400, <k>=10) (b) scale-free networks (N=400, <k>=10). (c) EEG networks (N=139, <k>=20).

Fig. 1. Activation exponents μ_s vs. attack exponents μ_k for a graph of N=400 and <k>=10, plotted for different values of randomization level p. The blue asterisks denote the exponents for scale-free and ER random graphs of same size and mean degree. Error bars indicate standard deviation. SRCC between μ_s and μ_k is r_s=0.99.



5 References

1. M. E. J. Newman, "The structure and function of complex networks," SIAM Rev., vol. 45, no. 2, pp. 167–256, Mar. 2003.
2. D. A. A. Stauffer, Introduction to percolation theory, 2nd ed. CRC press1994, .
3. B. Shrager, Jeff; Hogg, Tag; A. Huberman, "Observation of Phase Transitions in Spreading Activation Networks," Science (80-.)., vol. 236, pp. 1092 – 1094, 1987.
4. Erdős P. and Rényi A., "On the evolution of random graphs," Publ. Math. Inst. Hungar. Acad. Sci 5, pp. 17–61, 1960.
5. R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," Rev. Mod. Phys., vol. 74, no. 1, pp. 47–96, 2002.
6. J. Naim-Feil, M. Rubinson, D. Freche, A. Grinshpoon, A. Peled, E. Moses, and N. Levit-Binnun, "Altered brain network dynamics in schizophrenia: A cognitive-EEG study," Biol. Psychiatry CNNI, 2017.
7. B. C. M. van Wijk, C. J. Stam, and A. Daffertshofer, "Comparing brain networks of different size and connectivity density using graph theory.," PLoS One, vol. 5, no. 10, p. e13701, Jan. 2010.

Synchronized epidemic process and the possibly largest non-Markovian SIS threshold on networks

Qiang Liu and Piet Van Mieghem

Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands
Q.L.Liu@TUDelft.nl
P.F.A.VanMieghem@TUDelft.nl

1 Introduction. Since a real epidemic process [3] is not necessarily Markovian, the epidemic threshold obtained under the Markovian assumption may not be realistic. To understand general non-Markovian epidemic processes on networks, we study the Weibullian Susceptible-Infected-Susceptible (SIS) process in which the infection process is a renewal process [4] with a Weibull time distribution [1, 5]. The Markovian SIS process [6], which has been extensively studied, is a special case of the Weibullian SIS process.

In the Weibullian SIS process, two parameters matter. The first one is the well-known effective infection rate $\tau := \beta/\delta$, where β is the infection rate and δ is the Poissonian curing rate. The second parameter is the shape parameter α of the Weibull distribution, which controls the infection process. The Weibull distribution is heavy-tailed when $\alpha < 1$, exponential when $\alpha = 1$, hence Markovian, and Gaussian-like when $\alpha > 1$. Furthermore, tuning the shape parameter α dramatically shifts the epidemic threshold [5], and the epidemic threshold increases with the distribution changing from heavy-tailed to Gaussian-like (α increases). With the increase of α, the infection process becomes more synchronized. As shown in Fig. 1(a), the prevalence of the SIS process, which is the average fraction of infected nodes, on an Erdős-Rényi (ER) network has multiple peaks when $\alpha > 1$.

2 Results. The Weibullian SIS process with $\alpha \to \infty$ models the synchronized epidemic process where the infection happens exactly every $1/\beta$ time units. By a first-order mean-field approximation [1] similar to the N-Intertwined Mean-Field Approximation (NIMFA) for Markovian SIS processes [6], we obtain the epidemic threshold for $\alpha \to \infty$, which is $\tau_c^{(1)} := 1/\ln(\lambda_1 + 1)$, where λ_1 is the largest eigenvalue of the network's adjacency matrix. The threshold $\tau_c^{(1)}$ has a similar form as the NIMFA threshold $1/\lambda_1$ for Markovian SIS processes. If $\tau < \tau_c^{(1)}$ in the Weibullian SIS process with $\alpha \to \infty$, then the prevalence is upper bounded by an exponentially decreasing function with time t, while if $\tau > \tau_c^{(1)}$, then the process can enter a metastable state after enough long time, where the prevalence changes periodically, and the ratio between the maximum and minimum metastable prevalence cannot exceed $\lambda_1 + 1$ for any connected network. As shown in Fig. 1(b),1(c), and 1(d), our mean-field method approximates the exact process well. From the simulation results, the mean-field threshold $\tau_c^{(1)}$ seems to be, just as for NIMFA, a lower bound of the exact threshold.

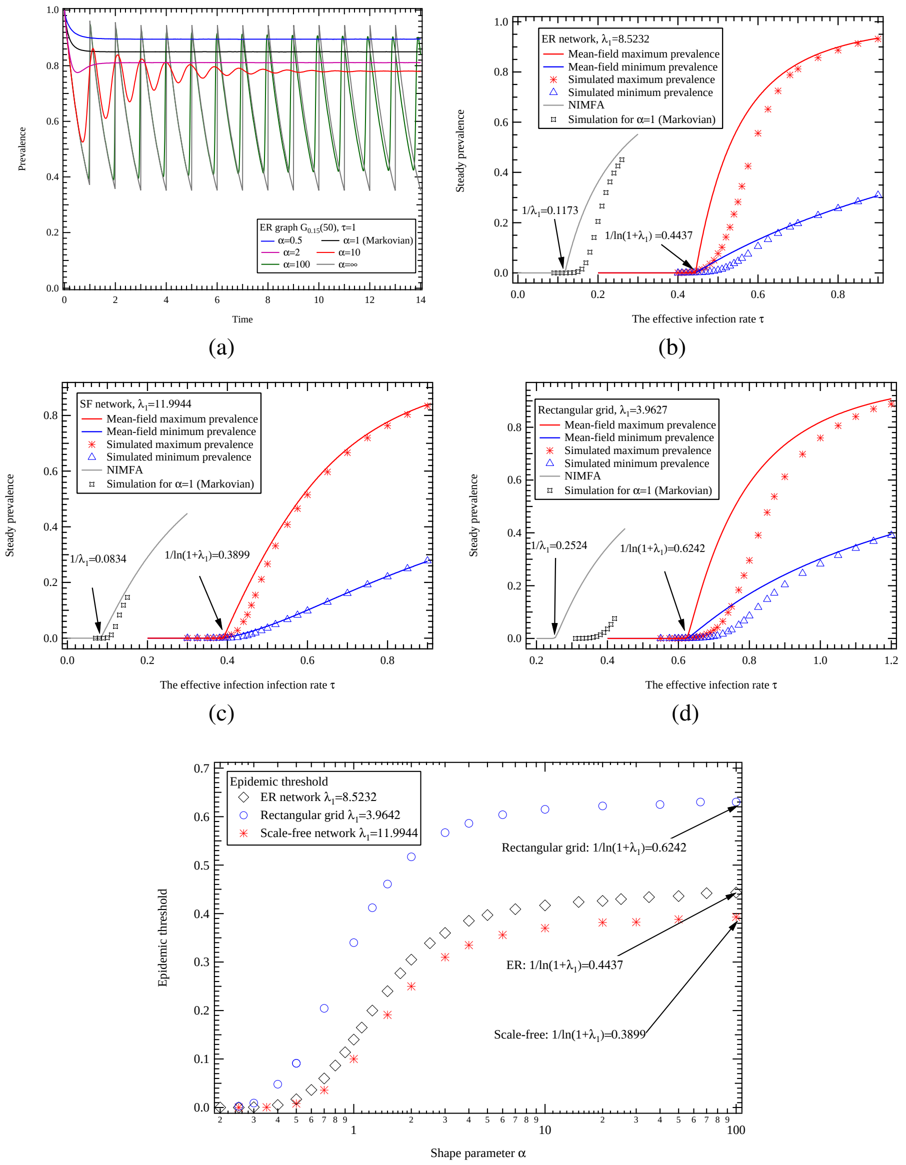
Since the Weibullian SIS process is capable of modeling various kinds of non-Markovian epidemic processes with a suitable value of α, $\tau_c^{(1)}$ is possibly the largest possible epidemic threshold of general non-Markovian SIS processes with a Poisson curing process under the mean-field approximation. If the effective infection rate $\tau > \tau_c^{(1)}$, then the infection will persist on the network for any finite $\alpha > 0$. When $\alpha \to 0$, no infected node can be cured [1] and the epidemic threshold is 0. Figure 1(e) shows the epidemic threshold for the different value of α, which is obtained by simulation of the exact Weibullian SIS process. The epidemic threshold increases approximately from 0 to $\tau_c^{(1)}$ with α.

3 Potential applications. The Weibullian SIS process with $\alpha \to \infty$ has a potential to model real situations. For example, many computer viruses burst periodically because the hackers spend time on improving the virus before each burst. Thus, the virus development life-cycle and the underlying network collectively determine whether the infection can persist or not. For another example, in a wireless sensor network, the clock of each sensor may need to be synchronized periodically because the clock will shift with time due to random noises. However, clock synchronization consumes computing and communication resources. Our results provide the maximum synchronization period which minimizes the resources consumption if the underlying network and the clock shifting rate are known. A third example is about advertising on social networks. Generally, an advertisement only has a short-time influence. After broadcasting, people discuss and spread the advertisement. However, the advertisement will be forgotten by the population in the long run, thus, the advertisement should be broadcasted periodically. Our results may help to find an optimal frequency of advertising.

By studying the Weibullian SIS process on networks, we model the synchronized SIS process and obtain a possibly largest possible epidemic threshold for SIS processes with a Poisson curing process. More details can be found in [1].

References

1. Liu, Q., Van Mieghem, P.: Burst of virus infection and a possibly largest epidemic threshold of non-markovian SIS processes on networks. Under review
2. Liu, Q., Van Mieghem, P.: Die-out probability in SIS epidemic processes on networks. In: International Workshop on Complex Networks and their Applications. pp. 511–521. Springer (2016)
3. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., Vespignani, A.: Epidemic processes in complex networks. Reviews of modern physics 87(3), 925 (2015)
4. Van Mieghem, P.: Performance analysis of complex networks and systems. Cambridge University Press, Cambridge (2014)
5. Van Mieghem, P., Van de Bovenkamp, R.: Non-Markovian infection spread dramatically alters the susceptible-infected-susceptible epidemic threshold in networks. Physical review letters 110(10), 108701 (2013)
6. Van Mieghem, P., Omic, J., Kooij, R.: Virus spread in networks. IEEE/ACM Transactions on Networking 17(1), 1–14 (Feb 2009)



(a)

(b)

(c)

(d)

(e)

Fig. 1. **(a)**: The prevalence of the Weibullian SIS process on an Erdős-Rényi (ER) network $G_{0.15}(50)$ obtained by averaging over 10^5 realizations. The effective infection rate $\tau = 1$ is around 8.5 times the NIMFA threshold ($1/\lambda_1 = 0.1173$). The minimum prevalence decreases with α. For $\alpha < 1$, the prevalence is a trivial single-peak function with time. For small $\alpha > 1$, the prevalence oscillates but eventually becomes approximately constant. When $\alpha \to \infty$, the metastable state prevalence is no longer constant. **(b), (c), (d)**: The metastable state maximum and minimum prevalence of three different networks under the mean-field approximation and simulation. The simulation runs for enough long time (50 time units with $\delta = 1$), and the maximum and minimum prevalence are selected from the last complete time period. The ER network is corresponding to Fig. 1(a). The Barabási-Albert scale-free network has 500 nodes and 1491 links. The rectangular grid has 484 nodes and 924 links. **(e)**: The epidemic threshold versus α. The threshold is chosen as the value of τ which leads to the maximum prevalence being around 0.001 at the last period in the simulation. All the simulation results are obtained by averaging over 10^5 realizations with all nodes infected initially (to prevent early die-out [2]).

Threshold driven contagion on weight heterogeneous networks

Samuel Unicomb[1], Gerardo Iñiguez[2,3], and Márton Karsai[1]

[1] Univ de Lyon, ENS de Lyon, INRIA, CNRS, UMR 5668, IXXI, 69364 Lyon, France
samuel.unicomb@ens-lyon.fr, marton.karsai@ens-lyon.fr
[2] IIMAS, Universidad Nacional Autonóma de México, 01000 Ciudad de México, Mexico
[3] Department of Computer Science, Aalto University School of Science, 00076 Aalto, Finland
gerardo.iniguez@aalto.fi

1 Introduction

Social influence is arguably among the main driving mechanisms of many collective phenomena in society, including the spreading of innovations, ideas, fads, or social movements. Many of these processes have been studied empirically in the past, particularly with regards to the existence of so-called adoption cascades, where large numbers of people adopt the same behaviour in a relatively short time. These phenomena have been commonly modelled either as simple contagion (where adoption is driven by independent contagion stimuli, like the Bass model of innovation diffusion [1]), or as complex contagion (where a threshold on the number of adopting neighbours in a social network determines spreading, like the Watts model of adoption cascades [2]). However, in these models social influence is usually considered homogeneous across ties in the network, implying that all acquaintances are equally likely to influence an ego while making decisions. In reality, the strength of social influence may vary from neighbour to neighbour as it depends on the intimacy, frequency, or purpose of interactions between acquaintances. Neglecting such local heterogeneities may lead to overly simplistic models and potentially undermine a detailed understanding of real spreading phenomena. We fill this gap by studying a simple yet realistic model of contagion, and find that diversity in interaction strength may radically speed up or slow down the process relative to its unweighted counterpart. Our results reveal the importance of weighted interactions for an accurate quantitative prediction of threshold process in nature and society.

2 Results

We introduce a dynamical cascade model on weighted networks, where tie heterogeneities capture diversity in social influence. We study our contagion model over synthetic and real weighted networks with computer simulations and approximate master equations (AME) [3], and explore the effect of model parameters in contagion for several weight distributions. First we focus on a bimodal weight distribution, such that spreading is determined by the adoption threshold ϕ of nodes (defined as the sum of link weights to adopting neighbours relative to the total strength of the actual node)

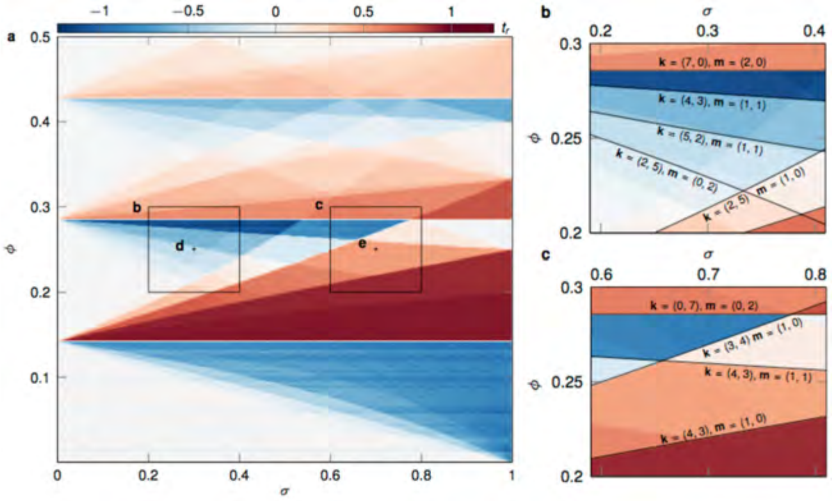


Fig. 1. Relative speed t_r of threshold driven cascades on weighted networks. (a) Relative time t_r of cascade emergence on (σ, ϕ)-parameter space, simulated over k-regular regular networks ($k = 7$) and averaged over 25 realizations. Time of cascades for given ϕ is either higher or lower than the corresponding case $(0, \phi)$ of an unweighted network. (b-c) Selected regions of parameter space in (a), where t_r is instead calculated from the numerical solution of the AME systems, while boundaries are obtained from a combinatorial arguments.

and the standard deviation σ of weight distribution. We find that the presence of tie weight heterogeneities induce unexpected dynamical behaviour, as they either speed up or slow down contagion with respect to the unweighed case, depending on ϕ and σ (Figure 1). We demonstrate this effect to be present in synthetic and data-driven simulations of adoption dynamics on various artificial and real networks. We show that the structure of this non-monotonous parameter space can be understood by combinatorial arguments, and we provide an analytical solution of the problem for networks with arbitrary degree and weight distributions, using approximate master equations [3]. These results [4] may be instrumental in developing more accurate spreading models that manage to gauge the rise and extent of real behavioural cascades in society.

References

1. Bass, F. M. A new product growth for model consumer durables. *Manage. Sci.* **15**, 215–227 (1969).
2. D. J. Watts, A simple model of global cascades on random networks, *Proc. Natl. Acad. Sci. U.S.A.* ***99**, 5766–5771 (2002).*
3. J. P. Gleeson, Binary-state dynamics on complex networks: Pair approximation and beyond. *Phys. Rev. X* **3**, 021004 (2013).
4. S. Unicomb, G. Iñiguez, M. Karsai, Threshold driven contagion on weighted networks. *arXiv:*1707.02185 (2017) (submitted).

Dimensionality and activity spreading in hierarchical modular networks

Ali Safari[1], Paolo Moretti[1], and Miguel Ángel Muñoz[2]

[1] Institute for Materials Simulation, Friedrich-Alexander University Erlangen-Nuremberg,
Dr-Mack-Str 77 90762 Fürth,Germany
ali.s.safari@fau.de

[2] Departamento de Electromagnetismo y Física de la Materia and Instituto Carlos I de Física
Teórica y Computacional
Universidad de Granada, Campus de Fuentenueva, E-18071 Granada, Spain

1 Introduction

Human brain networks [1,2], metabolic and regulatory networks [3,4] and fiber networks in connective tissue [5] are all well known examples of biological systems, which exhibit a hierarchical modular structure. Hierarchical modular networks (HMNs) are endowed with a finite topological dimension D, which naturally implies longer distances between network hotspots, as opposed to e.g. scale-free (SF) networks, where the largest hubs belong to close-by neighborhoods. Further singular aspects of HMNs are provided by their spectral properties [6]: i) HMNs exhibit small (although non-zero) spectral gaps; ii) HMNs exhibit eigenvector localization, not limited to the principal eigenvector. All such structural features suggest that activity spreading in HMNs may follow uncommon dynamic patterns. This conjecture is corroborated by the observation of Griffiths phases in HMNs, a dynamic signature of rare-region effects [6,7]. The open question remains, however, as to what mechanisms underlie the emergence of activity spreading – the epidemic threshold – in such network models. We addressed these issues in a recent work [8], whose results we will briefly summarize in this extended abstract.

In the following we will propose an extension of the well-known quenched mean-field (QMF) framework, which accounts for the spectral properties of HMNs. We will show that activity spreading in HMNs arises as a dynamic coalescence process, in which the unstable activation modes ascribed to several localized eigenvectors of the adjacency matrix $\mathbf{A}$ are able to mutually interact and lead to sustained activity. This picture is analogous to the re-infection mechanism recently proposed for networks with heavy-tailed degree distributions [9], even-though HMNs do not exhibit heavy-tailed degree distribution, nor do they contain large degree hubs. We will show that, although the principal eigenvalue alone cannot produce a correct estimate of the epidemic threshold in the spirit of the QMF result $\lambda_c^{QMF} = 1/\Lambda_1$, the topological dimension D provides and even stronger estimator, in the form $\lambda_c \sim 1/D$.

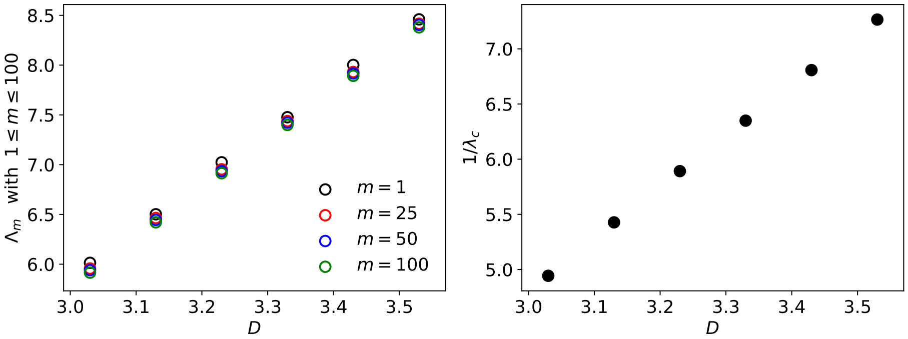


Fig. 1. Dependence of eigenvalues in the higher spectral edge (left) and epidemic thresholds (right) on the topological dimension D of HMNs. Results are numerical computations (left) and numerical simulations of SIS dynamics (right), in HMNs of size $N = 2^{20} \approx 10^6$.

2 Results

In order to make contact with the current literature, we model activity spreading as a standard SIS process, in which network nodes can be in either an active (infected) or inactive (susceptible) states, and activation (infection) is tuned by the spreading rate λ. At each time t, the probability that each node is active is given by the column vector $\boldsymbol{\rho}$, and the emergence of a stable active state $\boldsymbol{\rho}^\infty$ for any $\lambda > \lambda_c$ must comply with the well-known exact result [10]

$$\boldsymbol{\rho}^\infty \leq \lambda \sum_{m=1}^{N} \Lambda_m (\mathbf{v}_m \cdot \boldsymbol{\rho}^\infty)\mathbf{v}_m, \tag{1}$$

where Λ_m and $\mathbf{v}_m$ are the mth eigenvalue (sorted in descending order) and its corresponding eigenvector. In the standard QMF treatment, the spectral gap $\Lambda_1 - \Lambda_2$ is assumed to be large: this allows one to consider only the first term of the sum in Eq. (1), resulting in $\lambda_c^{QMF} = 1/\Lambda_1$ and $\boldsymbol{\rho}^\infty = \mathbf{v}_1$. As mentioned above, this approximation cannot be made in HNMs (as reflected by the fact that in HMNs $\lambda_c > 1/\Lambda_1$ [6]). We choose to relax this condition, by assuming that the critical state is not given by the principal eigenpair $(\Lambda_1, \mathbf{v}_1)$, but by the first m^* of them (in the worst case scenario, $m^* = N$ and all eigenpairs are needed). We find that a candidate active steady state $\boldsymbol{\sigma}^\infty$ is stable if it obeys

$$(\lambda\Lambda_m - 1)|\mathbf{v}_m \cdot \boldsymbol{\sigma}^\infty| \geq 0 \quad \forall m \leq m^*. \tag{2}$$

Eq. (2) suggests that an entire range of m^* eigenpairs may contribute to the epidemic threshold. This result is corroborated by our simulation results, which show that, upon varying HMNs realizations, the values of λ_c are not tied to any specific and/or pathological Λ_m (e.g. corresponding to a delocalized $\mathbf{v}_m$). Stably active states appear to emerge from the coalescing behavior of the individual short-lived states associated with each localized eigenvector. Can we still provide an estimate for λ_c, provided that a whole range (of unknown width) of eigenpairs is necessary? We show that the answer to this question is indeed affirmative. We demonstrate that all the eigenvalues in the upper spectral edge of a HMN scale with the topological dimension D, $\Lambda_m \sim D$ – not just the principal Λ_1, but also a wide range of lower eigenvalues Λ_m with $m > 1$ (Fig. 1, left). Since each corresponding eigenvector will contribute an unstable active state for a $\lambda \sim 1/\Lambda_m$, then the global and stable active state will obey the scaling law

$$\lambda_c \sim \frac{1}{D}. \tag{3}$$

Our simulation results (Fig. 1, right) confirm the prediction in Eq (3), finally providing us with an estimate for the epidemic threshold in HMNs.

Summary. We show that the onset of activity spreading (or epidemic threshold) in HMNs is the result of the complex dynamic interplay of a broad range of eigenvalues and eigenvectors of the adjacency matrix. This complex picture is however still tractable, as in HMNs all eigenvalues in the upper spectral edge are proportional to the topological dimension D, resulting in a remarkable scaling law that relates the epidemic threshold to the inverse of D.

References

1. Sporns, O., Tononi, G., Kötter, R.: The Human Connectome: A Structural Description of the Human Brain. PLoS Comput. Biol. 1(4), e42 (2005)
2. Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C J., Wedeen, V J., Sporns, O.: Mapping the structural core of human cerebral cortex. PLoS Biol. 6, e159 (2008)
3. Jeong, H., Albert, B T R., Oltvai, Z N., Barabási, A L.: The large-scale organization of metabolic networks. Nature 407, 651–654 (2000)
4. Ravasz, E., Somera, A L., Mongru, D A., Oltvai, Z N., Barabási, A L.:Hierarchical Organization of Modularity in Metabolic Networks. Science 297, 1551–1555 (2002)
5. Gautieri, A., Vesentini, S., Redaelli, A., Buehler, M J.: Hierarchical Structure and Nanomechanics of Collagen Microfibrils from the Atomistic Scale Up. Nano Lett. 11, 757–766 (2011)
6. Moretti, P., Muñoz, M A.: Griffiths phases and the stretching of criticality in brain networks. Nature Commun. 4, 2521 (2013)
7. Ódor, G., Dickman, R., Ódor, G.: Griffiths phases and localization in hierarchical modular networks Sci. Rep. 5, 14451 (2015)
8. Safari, A., Moretti, P., Muñoz, M. A.: Topological dimension tunes activity patterns in hierarchical modular networks. To appear in New Journal of Physics, https://doi.org/10.1088/1367-2630/aa823e (2017)
9. Boguñá M., Castellano C., Pastor-Satorras R.: Nature of the epidemic threshold for the susceptible-infected-susceptible dynamics in networks, Phys. Rev. Lett. 111, 068701 (2013)
10. Pastor-Satorras, R., Castellano C., Van Mieghem P., Vespignani A.: Epidemic processes in complex networks. Rev. Mod. Phys. 87, 925–979 (2015)

Fast diffusion of mutant mosquitoes in controlled evolutionary dynamics

Lorenzo Zino[1,2], Giacomo Como[2,3], and Fabio Fagnani[2]

[1] Department of Mathematics, University of Turin, 10129 Turin, Italy
lorenzo.zino@unito.it
[2] Department of Mathematical Sciences, Polytechnic University of Turin, 10123 Turin, Italy
fabio.fagnani@polito.it
[3] Department of Automatic Control, Lund University, 22100 Lund, Sweden
giacomo.como@control.lth.se

The study of diffusion processes on networks has improved our understanding in how ideas and innovations, as well as epidemics and mutations spread. The increased awareness of such spreading mechanisms paved the way for the design of effective control techniques, with inestimable potential benefits for the society.

In this work we focus on a hot topic in epidemics control. Mosquitoes are vehicle of several infections and diseases, such as Dengue fever, Malaria, and, recently, Zika. In the last few years, many efforts have been done by researchers to create harmless genetically modified organism (GMO), similar to the intermediate hosts, which can be introduced in nature [4, 3] to substitute the dangerous mosquitoes without modifying the environmental equilibria.

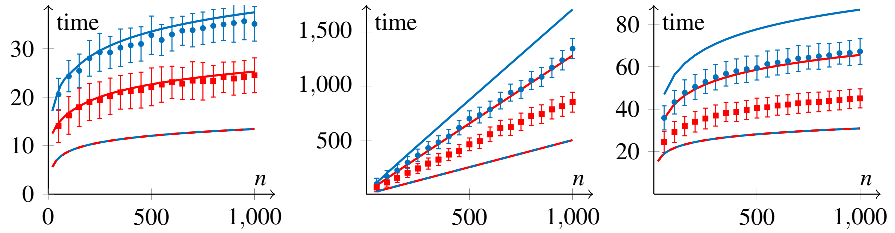
Inspired by the evolutionary dynamics [6], which have arisen as a powerful paradigm to study the spread of mutants in a geographic area, we propose a new model for mutant diffusion that presents two different features with respect to classical evolutionary dynamics: i) the diffusion process is modeled through link-based (instead of node-based as in most of the previous literature [6, 2, 1]) activation mechanisms; and ii) an explicit control action is incorporated. This change of perspective allows us to obtain, on the one hand, new analytical insights, far beyond the results available in the previous literature [6, 2], which are mainly based on extensive Monte Carlo simulations. On the other hand, it allows for the development of control policies to speed up the spread of the mutants. Specifically, we propose and analyze an effective feedback control strategy based on few knowledge on the network topology and on the evolution of the spreading process.

We model the geographic network as a connected weighted undirected graph $G = (V, E, W)$, where nodes are the locations and weighted links represent connections between locations. The systems evolution is driven by two factors: a *spreading mechanism* and an *external control*. The former acts as follows. Links activate according to Poisson processes (with rates from the weight matrix W), modeling the contact between mosquitoes occupying the two extremes of the link. If they belong to different species, then a conflict takes place and the winning species occupies both locations. Mutants win each conflict with a constant (independent) probability $\beta > 1/2$, modeling an evolutionary advantage of the GMOs. The spreading mechanism is thus an heterogeneous link-based biased voter model [7]. On the other hand, the external control consists in the introduction of mutants in a subset of nodes, called *target set*, modeling the procedure used to introduce GMOs in trials and case studies [4]. At first, we will consider the case in which both the target set and the rate of insertion in this set are set constant throughout the duration of the process.

Due to the external control, mutants eventually occupy the whole geographic area. The main question consists thus in understanding how the expected time needed for the mutants to invade the area is influenced by i) the network topology, and ii) the control policy adopted, though the choice of the target set and the rate of introduction. Here, we address this issue by presenting some analytical bounds on such quantity. Specifically, on the one hand, we propose a topology-based upper bound, related with the presence of bottlenecks in the network, on the other hand, we provide a pair of lower bounds in which the effect of the external control and the topology are strictly interlinked.

Using these analytical results, corroborated by Monte Carlo simulations, we identify families of graphs (such as expander graphs) where the mutants spread fast and occupy the whole network in a time growing logarithmically with the size of the network. On the other hand, we characterize other families of graphs yielding slow diffusion. Figs. 1(a-b) show the comparison between a fast diffusive topology (complete graphs) and a slow diffusive one (barbell graphs).

Finally, we discuss how the diffusion of the mutants can be speed up by designing a thoughtful time-varying control policy, in which the target set as well as the insertion rate can be dynamically changed. Specifically, we propose a feedback control policy that can strongly speed up the spreading process, guaranteeing fast diffusion in many situations in which the standard control policy fails, such as barbell graphs, as shown comparing Figs. 1(b-c). Beside its effectiveness, the strength of our control policy lies in its simplicity and feasibility. In fact, we let the target set to be a singleton and we move it in such a way that we always introduce mutants in locations occupied by the native species, without any optimization on that choice, which is known to be a computationally hard problem [5]. Moreover, we set the introduction rate as a feedback functions of only two macroscopic observables of the system.



(a) Complete graph (b) Barbell graph (c) Feedback control

Fig. 1: Monte Carlo estimation (200 simulations) with 90% confidence intervals of the expected absorbing time on different network topologies and increasing size of the network n and analytical bounds (solid lines). In red $\beta = 0.8$, in blue $\beta = 0.7$. In Figs.(a-b) the standard control policy is used, whereas in Fig.(c) is used a feedback control policy on barbell graphs.

Acknowledgment

This work was supported by the Swedish Research Council through Project Research Grant 2015-04066 and by the Compagnia di San Paolo through a PoliTo Starting Grant.

References

1. Allen, B., Lippner, G., Chen, Y.T., Fotouhi, B., Momeni, N., Yau, S.T., Nowak, M.A.: Evolutionary dynamics on any population structure. Nature 544(7649), 227–230 (2017)
2. Broom, M., Rychtář, J., Stadler, B.T.: Evolutionary dynamics on graphs - the effect of graph structure and initial placement on mutant spread. Journal of Statistical Theory and Practice 5(3), 369–381 (2011)
3. Carvalho, D.O., McKemey, A.R., Garziera, L., Lacroix, R., Donnelly, C.A., Alphey, L., Malavasi, A., Capurro, M.L.: Suppression of a field population of Aedes aegypti in Brazil by sustained release of transgenic male mosquitoes. PLoS Negl. Trop. Dis. 9(7), 1–15 (2015)
4. Harris, A.F., Nimmo, D., McKemey, A.R., Kelly, N., Scaife, S., Donnelly, C.A., Beech, C., Petrie, W.D., Alphey, L.: Field performance of engineered male mosquitoes. Nature Biotechnology 29(11), 1034–1037 (2011)
5. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03. p. 137. ACM Press, New York, New York, USA (2003)
6. Lieberman, E., Hauert, C., Nowak, M.A.: Evolutionary dynamics on graphs. Nature 433 (2005)
7. Ligget, T.M.: Interacting particle systems. Springer-Verlag, New York, NY, USA (1985)

Meta-foodwebs as a many-layer epidemic process on networks

Edmund Barter[1] and Thilo Gross[1]

University of Bristol
edmund.barter@brisotl.ac.uk,
www.biond.org

1 Introduction

Notable recent works have focused on the multi-layer properties of coevolving diseases. We point out that very similar systems play an important role in population ecology. Specifically we study a meta-foodweb model that was recently proposed by Pillai et al[**?**,**?**]. This model describes a network of species connected by feeding interactions, which disperse over a network of spatial patches, see Fig. **??**.

Focusing on the essential case, where the network of feeding interactions is a chain, we develop an analytical approach for the computation of the degree distributions of colonized spatial patches for the different species in the chain. This framework allows us to address ecologically relevant questions.

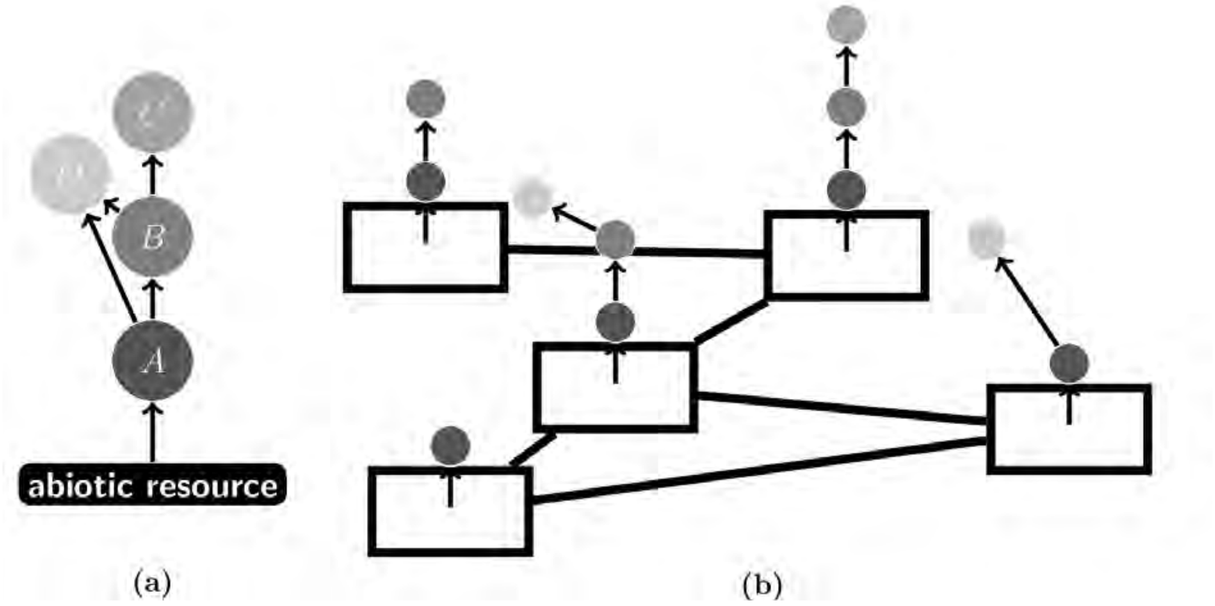


Fig. 1. Meta-foodwebs as networks of networks. Panel (a) shows a simple meta-foodweb of four species. Each node represents a species and the directed links are from prey to predators in a feeding relationship. Species *A* is a primary producer, while species *B* and *C* are specialist consumers. Species *O* is an omnivore, which can feed on multiple trophic levels. Panel (b) shows a patch network. Rectangles represent each patch and links represent potential dispersal routes for the species between the patches. Each patch is occupied by a local food chain. The local networks change in time due to colonization and extinction events.

2 Results

Considering configuration model ensembles of landscape networks, we find that the fraction of patches occupied by a species obeys

$$C \leq \frac{\langle k \rangle c}{4e}. \tag{1}$$

where the $\langle k \rangle$ is the mean degree, c is the colonisation rate and e is the extinction rate. Consequently there is an upper bound for the fraction of patches that a given species can occupy which depends only on the networks mean degree and the dispersal parameters.

For a given mean degree there is then an optimal degree distribution that comes closest to the upper bound. Notably scale-free degree distributions perform worse than more homogeneous degree distributions if the ration e/c is sufficiently low, see Fig. **??**.

The species in a foodchain experience the underlying network differently. Subsequently, the optimal degree distribution for one particular species is generally not the optimal distribution for the other species in the same food web.

These results are of interest for conservation ecology, where, for instance, the task of selecting areas of old-growth forest to preserve in an agricultural landscape, amounts to the design of a patch network.

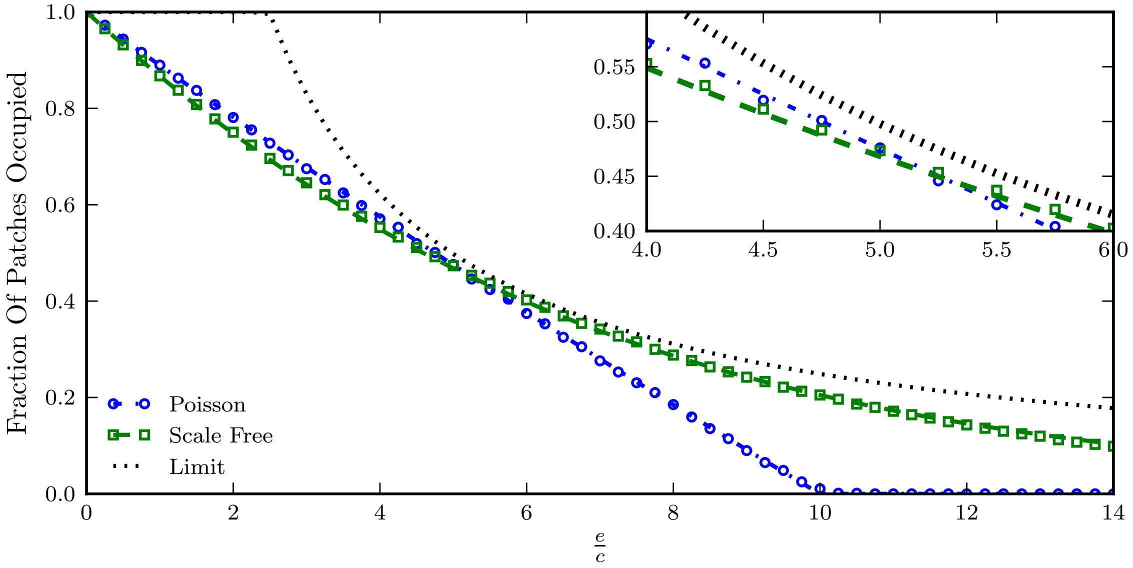


Fig. 2. Comparison of colonised abundance of species 1 in patch networks with different degree distributions. Shown are analytical (lines) and simulated (markers) abundances for species on patch networks with poisson (blue) and scale-free (green) degree distributions with the same mean degree and the limit as calculated from the configuration model. For low e/c the poisson degree distribution leads to greater abundance but for high e/c it is the scale-free distributions that allows greater abundance. The insert highlights the region at which the lines cross.

References

1. P. Pillai, M. Loreau, and A. Gonzalez, “A patch-dynamic framework for food web metacommunities,” *Theoretical Ecology*, vol. 3, pp. 223–237, dec 2009.
2. P. Pillai, A. Gonzalez, and M. Loreau, “Metacommunity theory explains the emergence of food web complexity.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. 19293–8, nov 2011.

Efficient network exploration via a core-biased random walk

Raúl J Mondragón

School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Rd, London E1 4NS, UK
r.j.mondragon@qmul.ac.uk

A simple strategy to explore a network is to use a random-walk where the 'random-walker' jumps from one node to an adjacent node at random. It is known that biasing the random jump, the walker can explore every walk of the same length with equal probability, this is known as a Maximal Entropy Random Walk (MERW) [1]. These biased random walks have been used to study problems in complex networks like link prediction [4] or the discovery of salient objects in an image [6]. To construct a MERW requires the largest eigenvalue and corresponding eigenvector of the adjacency matrix [1] which are global properties of the network. A good approximation to the MERW using only local properties is the degree-biased random walk [3, 2]. In this case the walker jumps from one node to a neighbouring node with a preference based on the degree of the destination node. In this note, we present a biased random walk where the walker prefers to jump to nodes that are in the core of the network.

The connectivity of a finite, undirected and connected network can be described by the symmetric adjacency matrix $\mathbf{A}$, where $a_{ij} = 1$ if nodes i and j share a link and zero otherwise. In these networks, a random walker would jump from node i to a neighbouring node j with a probability $P_{i\to j}$. The probability that the walker is in node j at time $t+1$ is $p_j(t+1) = \sum_i a_{ij} P_{i\to j} p_i(t)$ or in matrix notation $\overline{p}(t+1) = \pi \overline{p}(t)$. If the matrix π is primitive then the probability of finding the walker in node i as the times tends to infinity is given by the stationary distribution $\overline{p}^* = \{p_i^*\}$. In a network, $P_{i\to j}$ can be expressed as

$$P_{i\to j} = \frac{a_{ij} f_j}{\sum_j a_{ij} f_j} \tag{1}$$

where f_j is a function of one or several topological properties of the network, in this case the stationary distribution is [3]

$$p_i^* = \frac{f_i \sum_j a_{ij} f_j}{\sum_n f_n \sum_j a_{nj} f_j}. \tag{2}$$

The measure which tell us the minimum amount of information needed to describe the stochastic walk is the entropy rate $h = \lim_{t\to\infty} S_t/t$, where S_t is the Shannon entropy of all the walks of length t. This entropy is related to the properties of the random-walk via [1]

$$h = -\sum_{i,j} p_i^* P_{i\to j} \ln(P_{i\to j}). \tag{3}$$

The maximal entropy rate $h_{\max}$ corresponds to random walks where all the walks of the same length have equal probability. The value of $h_{\max}$ is related to the spectral properties of the network. If Λ is the largest eigenvalue and $\overline{v}$ its corresponding eigenvector of the adjacency matrix then the maximal entropy satisfies $h_{\max} = \ln(\Lambda)$ [1]. The MERW is obtained when the transition probability from node i to node j is $P_{i\to j} = a_{ij}v_j/(\sum_j a_{ij}v_j)$ [1], where v_i is the i–th entry of the eigenvector $\overline{v}$. The implementation of this biased random walk requires global knowledge of the network connectivity as we need to evaluate the largest eigenvalue-eigenvector pair. If this pair is not known, then a good approximation to the largest eigenvector $\overline{v}$ could be used to construct an approximation to the MERW. In a network where the nodes are ranked in decreasing order of their degrees, the connectivity of the network can be described with the degree sequence $\{k_r\}$ and the sequence of number of links $\{k_r^+\}$, where k_i^+ is the number of links that node i shares with nodes of higher rank. A bound to the largest eigenvalue in terms of the $\{k_r^+\}$ sequence is $\Lambda \geq 2\langle k^+\rangle_r$, where $\langle k^+\rangle_r = (1/r)\sum_i^r k_i^+$ is the average number of links shared by the top r ranked nodes [5].

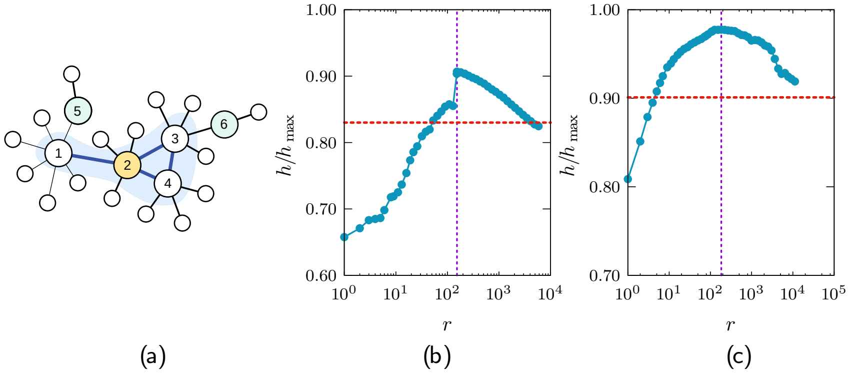


Fig. 1. (a) Simple network as an example of the core-biased jumps (see text). Ratio $h/h_{\max}$ for the (b) the co–authors network in High Energy Physics (HepTh) and (c) for the Autonomous System Internet.

Consider the adjacency matrix $\mathbf{A}$ of the network where the nodes are ranked in decreasing order of their degrees and $\overline{u}(r)$ a vector where its first r entries are set to one and the rest to zero. The vector $\overline{z}(r) = \mathbf{A}\overline{u}(r)$ has entries $z_i(r) = K_i^+(r)$, where $K_i^+(r)$ is the number of links that node i shares with the top r ranked nodes. Also notice that $\langle k^+\rangle_r = (\overline{u}^T(r)\,\mathbf{A}\,\overline{u}(r))/r$. As $2\langle k^+\rangle_r$ is a bound of Λ, then $\overline{z}(r_s)$ is an approximation to the eigenvector $\overline{v}$. A biased random jump based on $\overline{z}(r)$ is

$$P_{i\to j} = \frac{a_{ij}(K_j^+(r)+1)}{\sum_j^N a_{ij}(K_j^+(r)+1)}, \tag{4}$$

where the term 1 in the numerator and denominator was added as it is possible that $K_j^+(r) = 0$ if node j has no links with nodes of greater rank, and then the random walk will be ill defined. For the core-biased random walk the core is the value of r where the rate entropy is maximal, that is $r_c = \max_r(\{\arg\max_r h(r)\})$. The entropy rate $h(r)$ for this random walk is evaluated by considering $f_i = K_i^+(r) + 1$ in Eq. (2) to evaluate p_i^* and use this stationary probabilities in Eq. (3).

Figure 1(a) shows an example of the core-biased jumps, in the figure nodes 1 to 4 form the core. Nodes 5 and 6 both have degree equal to two, the jump probability from node 6 to node 3 is larger than the probability to jump from node 5 to node 1 because node 3 shares two links with the core against sharing only one link with the core for node 1. Similarly, node 2 which is inside the core, the probability to jump to node 3 or node 4 is greater than to jump to node 1, even that node 1 has larger degree, again this is because node 3 and 4 share more links with the core than node 1. Figure 1(b)-(c) shows the ratio $h/h_{\max}$ as a function of the core size r for the AS-Internet (disassortative) and the Hep-Th (assortative) networks. The vertical dashed line marks the value of r_c when h is maximal. The horizontal dashed line shows the value of $h/h_{\max}$ for the degree-biased random walk where $P_{i\to j} = (a_{ij}(k_j))/(\sum_j a_{ij}(k_j))$. The core size for these networks is small, the ratio of the size of the core against the number of node, r_c/N is 0.016 and 0.026 for the Internet and Hep-Th, respectively. We tested the core-biased against the degree-biased random walks for several real and synthetic networks and, in general, the core-biased outperforms the degree-biased random walk. The exception are regular networks where the core r_c is the whole network ($r_c = N$) and both methods give that $h = h_{\max}$ [3].

Finally, there is a large interval in r where the core-biased random walk performs better than the degree-biased random walk. This suggest that it is possible to create an efficient biased random walk even when the overall ranking of the nodes it is not known. If we assume that the core are the nodes with degree greater than k^* and $k^* < k_{\max}$ where $k_{\max}$ is the maximal degree of the network, then we can redefine $K_j^+(k^*)$ as the number of links that node j has with nodes of degree higher or equal to k^* and $P_{i\to j} = (a_{ij}(K_j^+(k^*) + 1))/(\sum_j a_{ij}(K_j^+(k^* + 1))$ could give an efficient biased random walk without having to evaluate the ranking of the nodes.

References

1. Burda, Z., Duda, J., Luck, J.M., Waclaw, B.: Localization of the maximal entropy random walk. Physical review letters 102(16), 160602 (2009)
2. Fronczak, A., Fronczak, P.: Biased random walks in complex networks: The role of local navigation rules. Physical Review E 80(1), 016107 (2009)
3. Gómez-Gardeñes, J., Latora, V.: Entropy rate of diffusion processes on complex networks. Physical Review E 78(6), 065102 (2008)
4. Li, R.H., Yu, J.X., Liu, J.: Link prediction: the power of maximal entropy random walk. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 1147–1156. ACM (2011)
5. Mondragón, R.J.: Network partition via a bound of the spectral radius. Journal of Complex Networks p. cnw029 (2016)
6. Yu, J.G., Zhao, J., Tian, J., Tan, Y.: Maximal entropy random walk for region-based visual saliency. IEEE transactions on cybernetics 44(9), 1661–1672 (2014)

Part V

Dynamics on/of Networks

Threshold cascade dynamics on signed networks

Kyu-Min Lee, Jae Woo Kim, and Kwang-Il Goh

Department of Physics, Korea University, Seoul 02841, Korea
E-mail: kgoh@korea.ac.kr

In many complex systems, not only do there exist positive interactions but also prevalent are the negative interactions between agents. Such systems can best be modeled as "signed" networks. There have been a body of research devoted in social network literature; however, quantitative understanding of the generic impact of such signed interactions on network dynamics is still lacking. Towards this goal here we study a generalized threshold cascade model originally due to Watts [3] on signed networks.

We define the activation and deactivation rule as follows. *Activation:* an inactive node gets activated if i) the fraction of active nodes among its neighbors in the positive-link layer exceeds the prescribed threshold R, and at the same time, ii) there is no active neighboring node in the negative-link layer. *Deactivation:* an active non-seed node can be deactivated if either the fraction of active neighbors in the positive layer drops to or below the threshold R or it encounters active neighbors in the negative layer.

The primary results of numerical simulations on random signed networks are displayed in Fig. 1. The presence of the negative interactions (encoded by the negative-link mean degree z_{neg}) is found effective in suppressing global cascade. The effect is most profound near the cascade boundary across which the global cascade disappears discontinuously, as is shown in the case of $z_{pos} \approx 4.0$ and $R = 0.2$. Notably, we found in this case the cascade size ρ exhibits phase transition-type behavior as the negative-link mean degree z_{neg} is increased, as manifested by the peak in the number of iterations in Fig. 1(f). Such peak is absent for R sufficiently away from there [Fig. 1(f)].

We also address using real-world signed network data the effect of the positive-negative degree correlations [1] to find their observable impact on cascade outcomes: The positive correlations between positive and negative layers suppresses the global cascades compared to its uncorrelated counterparts, while the negative correlations tend to promote the global cascades. Our model could readily be modified and generalized to be more detailed and realistic, including the variation in deactivation threshold or other cascade mechanisms such as the network coordination game-based cascade [2].

We anticipate this work could prompt the community to the study of dynamic processes on signed networks, which we believe will prove itself a fertile ground for yet another layer of complexity in network science.

References

1. Ciotti, V., Bianconi, G., Capocci, A., Colaiori, F., Panzarasa, P.: Degree correlations in signed social networks. Physica A 422, 25 (2015)
2. Shafaei, M., Jalili, M.: Community structure and information cascade in signed networks. New Gen. Comput. 32, 257 (2014)
3. Watts, D.J.: A simple model of global cascades on random networks. Proc. Natl. Acad. Sci. USA 99, 5766 (2002)

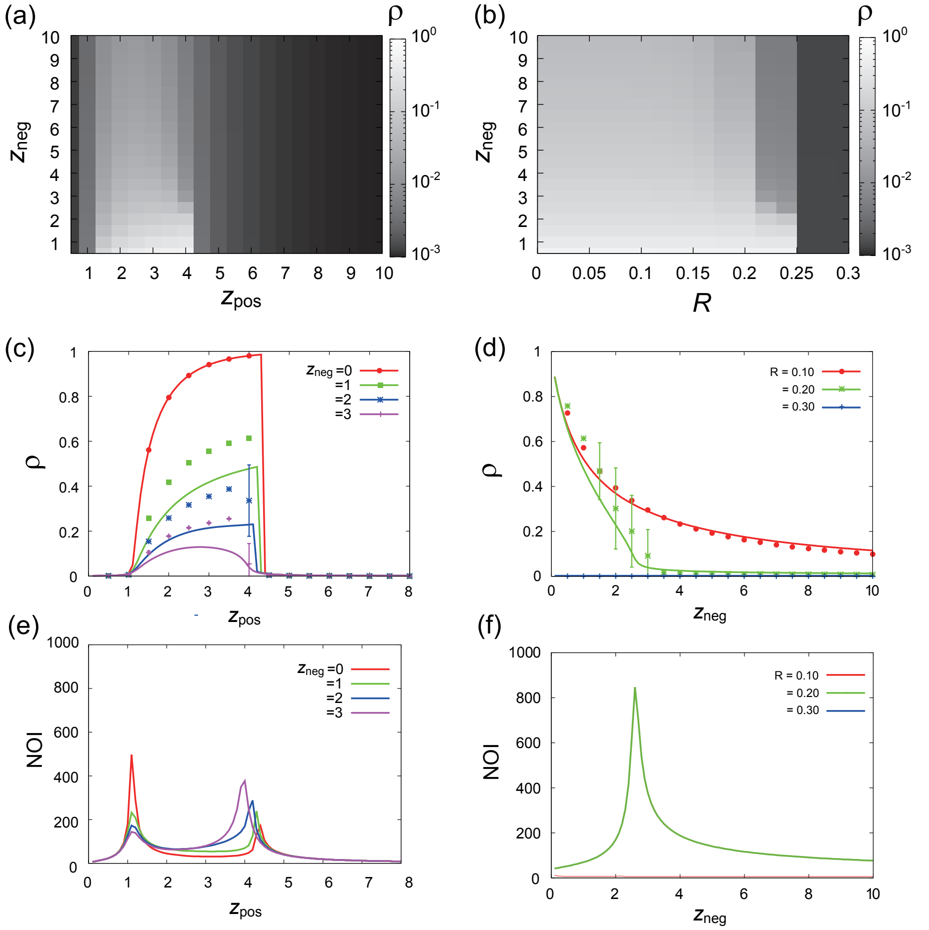


Fig. 1. Numerical simulation results on random signed networks. (a) The cascade size ρ as a function of the positive-link (z_{pos}) and negative-link (z_{neg}) mean degrees for fixed threshold $R = 0.2$. (b) The cascade size ρ as a function of the negative-link mean degree z_{neg} and the threshold R for fixed positive-link mean degree $z_{pos} = 4.0$. (c) The cascade size ρ is plotted against the positive-link mean degree z_{pos} for different values of the negative-link mean degree z_{neg} with fixed threshold $R = 0.2$. (d) The cascade size ρ is plotted against the negative-link mean degree z_{neg} for different threshold R with fixed positive-link mean degree $z_{pos} = 4.0$. (e, f) The number of iterations (NOI) required to reach stationary states is plotted for the cases of (c, d). All numerical results are from simulations on uncorrelated two-layer random networks, with each layer being Erdős-Rényi network of positive and negative links respectively, of size $N = 10^5$. The initial seeds of fraction $\rho_0 = 10^{-3}$ are randomly placed.

Invariant Collective Dynamics Under Network Transformations

Lluís Arola-Fernández[1] and Alex Arenas[1]

Dep. d'Eng. Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona, 43007, Spain
lluis.arolaf@urv.cat

The study of dynamical processes on complex networks has received a large amount of research due to the wide range of applications in many real systems and the increase in availability and precision of empirical data. It is possible to measure the dynamic response of many complex systems, but usually there is little available information about the network topology, its evolution and the local interaction mechanisms [1]. Under these circumstances, an accurate prediction of the behavior of the system may be still out reach, but it is already possible to tackle the problems of inference [2] and optimization [3] of the network configuration for an observable dynamic response. This is crucial for practical applications, specially in the design of optimal protocols to transform a network topology to achieve a desired performance or to reconstruct the microscopic configuration when only partial information is available from measurements.

Inspired by the derivation of Statistical Mechanics from Information Theory as a particular case of statistical inference [4], we propose a new methodology to find the network transformations that preserve the observable macro-state. We impose a generalized mean-field constraint in the transformation (conservation of node's strength), and optimize the distribution of weights to reconstruct the microscopic configuration using the available information, obtaining mapping protocols to construct functionally equivalent networks even when they have completely different connectivity patterns.

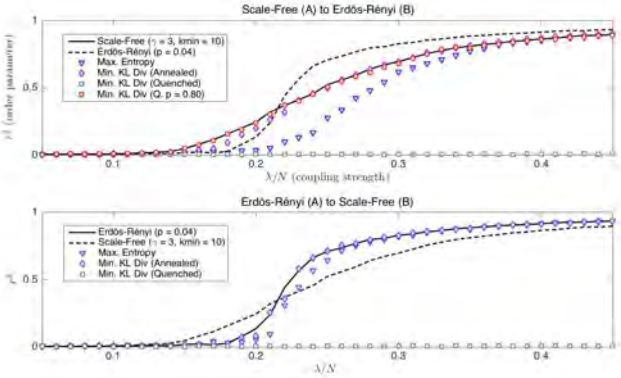


Fig. 1. Average steady-state $\langle r^2 \rangle$ depending on λ (coupling strength) for the unweigthed SF (solid line), ER (dashed line) and the transformed -weighted- ER networks (markers) using different information-theoretic tools: Principle of Maximum Entropy and Minimum KL Divergence.

Figure 1 shows the macroscopic evolution of the system in the Kuramoto Model, the most celebrated approach to describe the synchronization process of phase oscillators [5]. Results are shown for a particular scenario: an Erdos-Renyi network that transforms into an Scale-Free network (with exponent $\gamma = 3$), while preserving the number of nodes and density of connections $(N = 10^3, p \simeq 0.04)$. Several transformations, with different states of information available, are applied to analytically tune the weights of the connections in the ER topology in order to achieve accurate reconstructions of the collective behavior that emerges from the unweighted SF network.

In our work [6], we expose the assumptions of the method, the derivation of the analytical transformations and we also present numerical results for other dynamical processes: the SIS Model for epidemic spreading and the Ising model. Due to the simplicity and abstraction of the formalism, it turns out to be applicable to the analysis of systems of coupled individuals with arbitrary interaction mechanisms, thus providing a general tool to work with dynamical processes on networks when dealing with uncertainty in the measurements.

References

1. M.E.J. Newman, Networks, An Introduction, Oxford University Press (2010).
2. M. Timme, Revealing network connectivity from dynamic response, PRL 98 (2007).
3. C. Zhou, A.E. Motter, J. Kurths, Universality in the synchronization of weighted random networks, PRL. 96, 034101 (2006)
4. Jaynes, E.T. Information theory and statistical mechanics. Phys.Rev. 106 (4), (1957).
5. A. Arenas et al., Synchronization in complex networks, Phys. Rep. 469, 93 (2008).
6. In preparation.

Generic features of temporal evolution in hierarchical complex systems

Gerardo Iñiguez[1,2], Carlos Pineda[3], Carlos Gershenson[1], Sergio Sánchez[3], José A. Morales[3], and Albert-László Barabási[4,5,6]

[1] IIMAS, Universidad Nacional Autonóma de México, 01000 Ciudad de México, Mexico
[2] Department of Computer Science, Aalto University School of Science, 00076 Aalto, Finland
gerardo.iniguez@aalto.fi
[3] Instituto de Física, Universidad Nacional Autonóma de México, 01000 Ciudad de México, Mexico
[4] Center for Complex Network Research, Northeastern University, 02115 Boston, MA, USA
[5] Center for Cancer Systems Biology, Dana-Farber Cancer Institute, 02115 Boston, MA, USA
[6] Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 02115 Boston, MA, USA

Significance

Many complex systems are hierarchical in nature; social groups, economic and biological ecosystems, transportation infrastructures, languages, they all develop hierarchies of their constitutive elements that emerge from networked interactions within the system and with the external world. These hierarchies change in time according to system-dependent mechanisms of interaction, such as selection in evolutionary biology, or rules of performance in human sports, and reflect the relevance or ability of the element in performing a function in the system. However, it is still unclear whether the temporal evolution of hierarchies solely depends on the driving forces and characteristics of each system, or if there are generic features of hierarchy stability that allow us to model and predict patterns of hierarchical behaviour without considering the particularities of the system [1]. We explore this question by analysing over 30 datasets of social, nature, economic, infrastructure, and sports systems in a wide range of sizes ($10^2 - 10^5$) and time scales (from days to centuries) and find that, despite their various origins, the elements in these systems show remarkably similar stability depending on their position in the hierarchy. By classifying systems from closed to increasingly open, we manage to reproduce their hierarchy evolution in a minimal model with no system-dependent mechanisms of interaction. This allows us to make predictions on unobserved data, such as the likelihood of an unknown element climbing high in the hierarchy, or the time scale over which an element can maintain its relevance in the system. Our results may be crucial in further understanding why hierarchies evolve similarly in seemingly unrelated areas, and give clues on how to promote stability in the complex socio-technical systems of our day.

Methods

We consider hierarchy in these empirical systems as an ordered set of N elements ranked in relevance from most (rank $R = 1$) to least ($R = N$) relevant according to some prop-

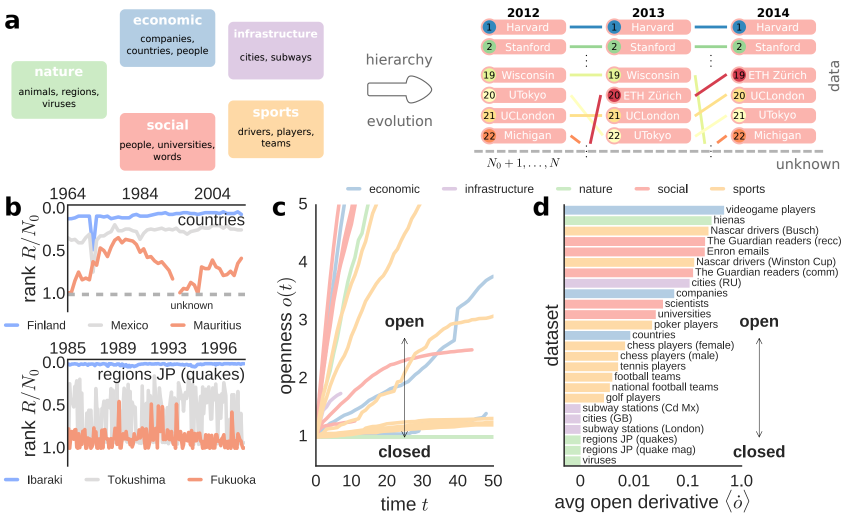


Fig. 1. Generic features of temporal evolution in hierarchical complex systems. **(a)** Datasets of social, nature, economic, infrastructure, and sports systems in a wide range of sizes and time scales, where data corresponds to the N_0 (out of N) most relevant elements in the system across time, such as the top universities in the world year by year. **(b)** Temporal evolution of rank R for elements in typical closed (bottom) and open (top) systems. Closed systems (like regions in Japan ranked by number of earthquakes) are symmetric in the sense that both highly and lowly ranked elements are stable across time, while in open systems (like countries ranked by economic complexity) the least relevant elements show the most fluctuations in rank. **(c-d)** System openness $o(t)$ as a function of time t (c), and its average derivative $\langle \dot{o} \rangle$ (d) for all datasets. Social, economic and sports systems tend to be open, while some nature and infrastructure systems are closed.

erty. We measure how elements change ranks across time, but we only have data on the N_0 most relevant ranks, such as top universities, most developed countries, or regions with the most earthquakes (Fig. 1a-b). We then measure the openness $o(t)$ as the number of elements that have visited ranks $R = 1,\dots,N_0$ relative to N_0 up until time t, and classify systems from closed to increasingly open based on the temporal behaviour of this quantity (Fig. 1c-d). Closed systems are symmetric in the sense that both highly and lowly ranked elements are stable across time, while in open systems the least relevant elements show the most fluctuations in rank. We emulate these generic features in a minimal model where elements change ranks randomly, and fit two unknown properties of each system: the size N and a microscopic time scale of hierarchy evolution. Finally, we validate the model with several measures (such as rank diversity [2, 3] and the probability that a rank changes elements in time) and make predictions on the future behaviour of the system: the probability that an element with unknown rank (larger than N_0) will become more relevant, and the typical time (beyond current observations) an element in ranks $R = 1,\dots,N_0$ will remain there.

References

1. Blumm, N., Ghoshal, G., Forró, Z., Schich, M., Bianconi, G., Bouchaud, J.-P., Barabási, A.-L.: Dynamics of ranking processes in complex systems. Phys. Rev. Lett. 109, 128701 (2012).
2. Cocho, G., Flores, J., Gershenson, C., Pineda, C., Sánchez, S.: Rank diversity of languages: Generic behavior in computational linguistics. PloS ONE 10, e0121898 (2015).
3. Morales, J. A., Sánchez, S., Flores, J., Pineda, C., Gershenson, C., Cocho, G., Zizumbo, J., Rodríguez, R. F., Iñiguez, G.: Generic temporal features of performance rankings in sports and games. EPJ Data Science 5, 33 (2016).

Stochastic textual block modelling in dynamic networks

Marco Corneli[1], Charles Bouveyron[2], Pierre Latouche[1] and Fabrice Rossi[1]

[1] Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne.
E-mail: Marco.Corneli@malix.univ-paris1.fr
[2] Laboratoire J.A. Dieudonné, UMR CNRS 7351 Equipe Asclepios, INRIA Sophia-Antipolis
Université Côte d'Azur, Nice, France

1 Outline

Social networks describe interactions between actors. In most cases, these interactions occur through document exchanges. Unfortunately, most clustering methods in network analysis cannot analyze the text contents and only focus on the"who talks to whom" information. The stochastic topic block model (STBM, [2]) was proposed to tackle this issue. It generalizes both SBM (stochastic block model, [6]) and LDA (latent Dirichlet allocation, [1]). The basic principles of STBM are summarized in the following steps: i) interactions between nodes are generated according to SBM, ii) an interaction corresponds to a document sent from one node i to another node j, iii) the proportion of topics discussed in the document only depends on the clusters of i and j. Hence, STBM seeks to detect node groups that are homogeneous both in terms of connection profiles and discussed topics. We propose an extension of STBM, called dSTBM, dealing with dynamic graphs (e.g. sequences of graphs snapshots). In order to avoid possible identifiability issues ([4]), in our model node clusters are fixed in time and we look for changes in the way the existing clusters interact with each other. Given M nodes, an hidden vector Y of length M is introduced such that $Y_i = q$ if node i belongs to the q-th cluster. Similarly, given U snapshots, an hidden vector X of length U is introduced such that $X_u = l$ if the u-th snapshot is drawn from the l-th time cluster. Using an equivalent 0-1 notation $Y_{iq} = 1$ iff $Y_i = q$, etc. A crucial assumption in dSTBM is

$$D_{iju}|Y_{iq}Y_{jr}X_{ul} = 1 \sim \mathscr{P}(D_{iju};\lambda_{qrl}),$$

where D_{iju} is the number of documents sent from i to j in the u-th snapshot and $\mathscr{P}(\cdot;\lambda)$ denotes a Poisson probability distribution of parameter λ. This assumption states that the *expected* number of documents sent from i to j in the u-th snapshot only depends on the clusters of i and j and on the time cluster of the u-th snapshot. The words in these documents follow a mixture distribution over K latent topics. The corresponding vector of topic proportions has length K and it is denoted by θ_{qrl}. It also depends on the clusters of i and j and on the time cluster of the u-th snapshot. We developed an inference procedure to estimate Y, X and θ. A model selection criterion was derived also to estimate the number of clusters (Q), time clusters (L) and topics (K).

1.1 Related works

Among probabilistic methods for text analysis, the latent Dirichlet allocation (LDA, [1]) is quite popular. The basic idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The topic proportions are assumed to follow a Dirichlet distribution. The author-topic (AT, [10][8]) and the author-recipient-topic (ART, [5]) models partially extend LDA to deal with textual networks. Although providing authorships and information about recipients, these models do not account for the graph structure, e.g. the way vertices are connected. A first attempt to take into account the graph structure, along with the textual content of edges is due to [11]. The authors propose two community-user topic (CUT) models: CUT1, modeling the communities based on the graph structure only and the CUT2, modeling the communities based on the textual information alone. More recently, [7] extended the ART model by introducing the community-author-recipient-topic (CART) model. In this context, authors and recipients are assigned to latent communities and they are clustered by CART based on homogeneity criteria, both in terms of graph structure and textual content. Interestingly, the nodes are allowed to belong to multiple communities and each pair of nodes is associated with a specific topic. Although flexible, the models illustrated so far rely on Gibbs sampling for the inference procedure, which can be prohibitive when dealing with large networks. An alternative model, that can be fitted via variational EM inference, is the topic-link LDA ([3]) performing both community detection and topic modeling. This model employs a logistic transformation based on topic proportions as well as author latent features. A family of 4 topic-user-community models was proposed by [9]. These models, accounting for multiple community/topic memberships, discover topic-meaningful communities in graphs with different types of edges. This is of particular interest in social networks like Twitter where different types of interactions exist: follow, tweet, re-tweet, etc.

2 The dSTBM at work: the Enron dataset

The dynamic network motivating this work is the Enron data set, containing all the e-mail communications between 149 employees of the company, from 1999 to 2001. The whole year 2001 is selected as the time horizon and partitioned in sub-periods (21 weeks). We aggregated interactions/e-mails to obtain a sequence of 21 static graphs whose edges are associated with the exchanged e-mails over the corresponding weeks. The considered time window spans from September, 3rd, 2001 to January, 28th, 2002, including three key dates i) September, 11th, 2001: the terrorist attacks to the Twin Towers and the Pentagon (USA). ii) October, 31st, 2001: the Securities and Exchange Commission (SEC) opened an investigation for fraud concerning Enron. iii) December, 2nd, 2001: Enron failed for bankruptcy (more than 4,000 lost jobs). The model selected *nine* topics, *six* node groups and *four* time segments. In Figure 1, an histogram reports the frequency of exchanged e-mails in the whole graph, each rectangle covering one week. Rectangles/weeks of the same color are assigned to the same time cluster by dSTBM. Notice that the time cluster changes occur some days after the three key dates mentioned above, represented in the figure by three vertical lines: black, blue and red, respectively. Figure 2 shows four graph snapshots of the Enron dataset. Each snapshot is obtained by aggregating the interactions/e-mails over the corresponding time cluster. Vertices of the same color are assigned to the same cluster by the inference algorithm

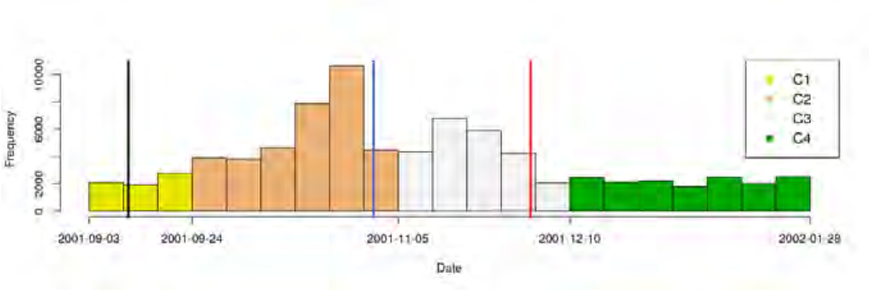
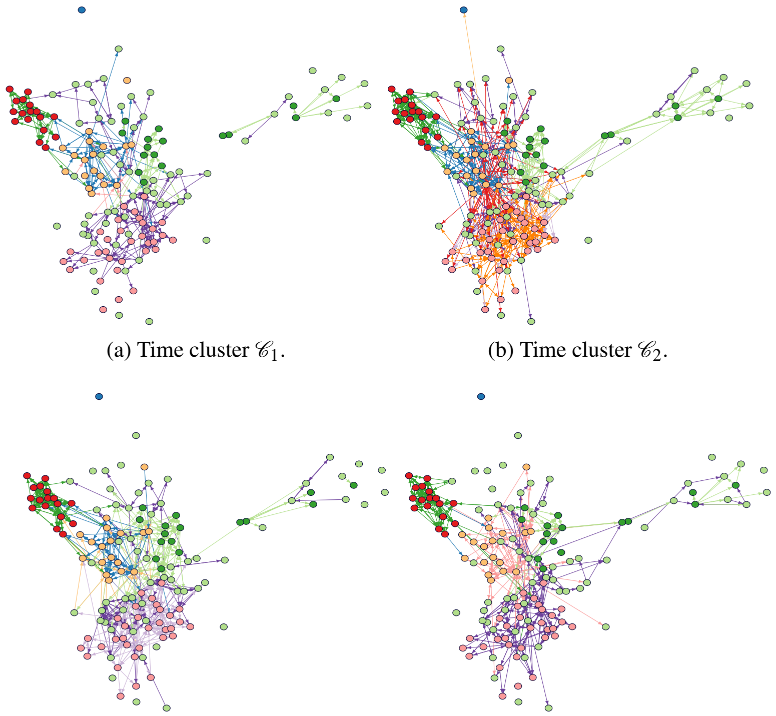


Fig. 1: Time clustering results obtained by dSTBM for the Enron data set (Sept. 2001 - Jan. 2002). The vertical lines mark: the 9/11 attacks (black), the investigation opening by SEC (blue), the company failure (red).

and edges of the same color are associated with the same main topic on the considered time segment. Interestingly, in the second snapshot, the topic associated with the orange edges contains words like “afghanistan” and “taleban” and it is clearly related to Enron activities in Afghanistan: Enron and the Bush administration were suspected to work secretly with Talebans before the 9/11 attacks. This topic appears in the graph during the second time segment starting on September, 24th, 2001, exactly two weeks after the 9/11 attacks. The “orange” topic disappears in later time clusters.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (Mar 2003), http://dl.acm.org/citation.cfm?id=944919.944937
2. Bouveyron, C., Latouche, P., Zreik, R.: The stochastic topic block model for the clustering of vertices in networks with textual edges. Statistics and Computing (2016), https://hal.archives-ouvertes.fr/hal-01299161
3. Liu, Y., Niculescu-Mizil, A., Gryc, W.: Topic-link lda: Joint models of topic and author community. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 665–672. ICML ’09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1553374.1553460
4. Matias, C., Miele, V.: Statistical clustering of temporal networks through a dynamic stochastic block model. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79(4), 1119–1141 (2017)
5. McCallum, A., Corrada-Emmanuel, A., Wang, X.: The author-recipient-topic model for topic and role discovery in social networks. In: Workshop on Link Analysis, Counterterrorism and Security (2005)
6. Nowicki, K., Snijders, T.: Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association 96(455), 1077–1087 (2001)
7. Pathak, N., DeLong, C., Banerjee, A., Erickson, K.: Social topic models for community extraction (2008)
8. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. pp. 487–494. UAI ’04, AUAI Press, Arlington, Virginia, United States (2004), http://dl.acm.org/citation.cfm?id=1036843.1036902



(a) Time cluster $\mathscr{C}_1$.

(b) Time cluster $\mathscr{C}_2$.

(c) Time cluster $\mathscr{C}_3$.

(d) Time cluster $\mathscr{C}_4$.

Fig. 2: Clustering results obtained by dSTBM for the Enron data set (Sept. 2001 - Jan. 2002). Each graph corresponds to a time cluster.

9. Sachan, M., Contractor, D., Faruquie, T.A., Subramaniam, L.V.: Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st International Conference on World Wide Web. pp. 331–340. WWW '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2187836.2187882
10. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 306–315. KDD '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/1014052.1014087
11. Zhou, D., Manavoglu, E., Li, J., Giles, C.L., Zha, H.: Probabilistic models for discovering e-communities. In: Proceedings of the 15th International Conference on World Wide Web. pp. 173–182. WWW '06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1135777.1135807

Pair approximation for the q-voter model with independence on complex networks

Arkadiusz Jędrzejewski

Department of Theoretical Physics, Faculty of Fundamental Problems of Technology, Wrocław University of Science and Technology, Wrocław, Poland,
arkadiusz.jedrzejewski@pwr.edu.pl

1 Introduction

In recent years, especially considerable attention has been drawn to the q-voter model [1–9], which was originally proposed by Castellano *et al.* [10] and later altered by Nyczka *et al.* [11]. Its modified version represents opinion dynamics under two types of social response: conformity and independence [12]. Although widely exploited in sociophysics, it also finds application in computational economics as an underlying mechanism of consumers decision making process, sometimes with slight modifications [13–15].

Although researchers associate the q-voter model with opinion dynamics, it is frequently allocated on simple topologies like complete graphs or regular lattices [4–7] despite the fact that social networks are regarded as complex structures. And even if some studies consider irregularities and randomness in modeling pairwise relations between agents, they focus primarily on simulation results, and only minority proposes analytical approach, as well [16–18]. Thus, in this paper, we carry out a mathematical analysis into the problem, and we clearly establish a connection between q-voter dynamics and complex networks. The system is studied by using pair approximation, which was previously employed only to the linear voter model [17] and has not been applied before to the nonlinear one. Moreover, the paper refers to the q-voter dynamics with stochastic noise, sociologically interpreted as independence, for which analytical results are known only in the case of a complete graph [11]. The obtained approximation is validated by carrying out Monte Carlo simulations on several structures like random regular (RR), Erdős-Rényi (ER), Watts-Strogatz (WS), and scale-free (SF) networks, including the Barabási-Albert (BA) model. We also show that in the limiting case our prediction coincides with the solution for a complete graph.
The study was published in Ref. [19].

2 Model description

We consider an arbitrary network of the size N. Each vertex is associated with one autonomous agent characterized by a binary, spin like variable $s_i = \pm 1$. In every elementary time step, we pick at random an agent and a group of influence comprised of q its randomly selected neighbors. The group is called q-panel, and its potential members are determined by the network topology and selected at random. With probability p, the chosen agent acts independently and adopts the opposite opinion or preserves the old one with equal chances. Otherwise, with probability $1-p$, it behaves like a conformist and embraces the viewpoint of q-panel, but only if the group is unanimous, that is to say, all q individuals have the same state. Drawing agents occurs without repetition so that the above definition of the model is identical to that in Ref. [11]. Nevertheless, now we extend underlying topologies to more complex structures than complete graphs.

3 Results

We were mainly interested in the time evolution and stationary values of the up-spin concentration. It turns out that the qualitative behavior of a system on studied weakly clustered complex networks is similar as on a complete graph and depends on the model parameter q, that is to say, for $q \leq 5$, the system undergoes continuous phase transitions, whereas for $q \geq 6$, phase transitions are discontinuous. However, the quantitative behavior also depends on the average node degree of an underlying network $\langle k \rangle$. What is interesting is that networks which have very different arrangements of edges and node degree distributions lead to the same results when they have the same value of the average node degree. Based on the pair approximation, we can derive formulas for the phase diagrams and the critical value of p in the case of continuous phase transitions. Along with increasing value of $\langle k \rangle$, the phase diagram shifts toward higher values of independence p; see Fig. 1. Consequently, it results in higher critical point. Moreover, for large average degrees we showed that solutions established from the pair approximation converge to those from the mean-field theory.

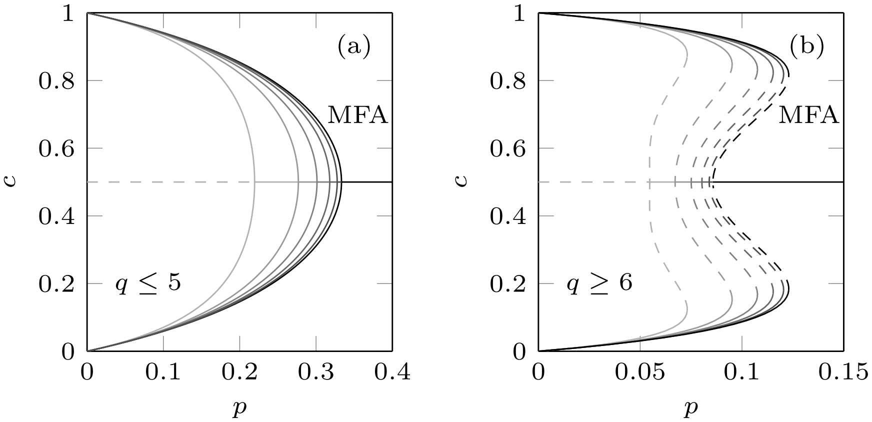


Fig. 1. A schematic representation of the pair approximation for the q-voter dynamics with independence. Darker lines correspond to the higher average degree of a network. The last, black line is the outcome of the mean-field approximation (MFA). The model exhibits two types of phase transitions: (a) continuous and (b) discontinuous.

Acknowledgments

The work was partially supported by funds from the National Science Centre (NCN, Poland) through Grant No. 2013/11/B/HS4/01061. The author thanks Katarzyna Sznajd-Weron for scientific guidance, meaningful discussions, and suggestions.

References

1. M. A. Javarone and T. Squartini, J. Stat. Mech.: Theory Exp. 2015, P10002 (2015).
2. A. Jędrzejewski, K. Sznajd-Weron, and J. Szwabiński, Physica A 446, 110 (2016).
3. P. Siedlecki, J. Szwabiński, and T. Weron, J. Artif. Soc. Soc. Simulat. 19, 9 (2016).
4. A. Mellor, M. Mobilia, and R. K. P. Zia, Europhys. Lett. 113, 48001 (2016).
5. M. Mobilia, Phys. Rev. E 92, 012803 (2015).
6. A. M. Timpanaro and C. P. C. Prado, Phys. Rev. E 89, 052808 (2014).
7. A. M. Timpanaro and S. Galam, Phys. Rev. E 92, 012807 (2015).
8. A. Chmiel and K. Sznajd-Weron, Phys. Rev. E 92, 052812 (2015).
9. K. Sznajd-Weron and K. M. Suszczynski, J. Stat. Mech.: Theory Exp. 2014, P07018 (2014).
10. C. Castellano, M. A. Muñoz, and R. Pastor-Satorras, Phys. Rev. E 80, 041129 (2009).
11. P. Nyczka, K. Sznajd-Weron, and J. Cisło, Phys. Rev. E 86, 011105 (2012).
12. P. R. Nail and K. Sznajd-Weron, Acta Phys. Pol., A 129, 1050 (2016).
13. A. Kowalska-Pyzalska, K. Ćwik, A. Jędrzejewski, and K. Sznajd-Weron, Acta Phys. Pol., A 129, 1055 (2016).
14. K. Maciejowska, A. Jędrzejewski, A. Kowalska-Pyzalska, and R. Weron, Acta Phys. Pol., A 129, 1045 (2016).
15. K. Byrka, A. Jędrzejewski, K. Sznajd-Weron, and R. Weron, Renewable Sustainable Energy Rev. 62, 723 (2016).
16. P. Moretti, S. Liu, C. Castellano, and R. PastorSatorras, J. Stat. Phys. 151, 113 (2013).
17. F. Vazquez and V. M. Eguíluz, New J. Phys. 10, 063011 (2008).
18. V. Sood, T. Antal, and S. Redner, Phys. Rev. E 77, 041121 (2008).
19. A. Jędrzejewski, Phys. Rev. E 95, 012307 (2017).

Interplay of time scales on the dynamics of complex networks

Kajari Gupta and G. Ambika

Indian Institute of Science Education and Research, Pune-411008, India
kajarig@students.iiserpune.ac.in,
g.ambika@iiserpune.ac.in

1 Introduction

Many complex systems that occur in physical, biological, chemical and geophysical contexts have large number of sub systems or units interacting with each other that may have different dynamical time scales[1–4]. Here, we study the interplay between time scales of dynamical systems that are connected on a complex network, where out of N systems m evolve on a slower time scale. Using the framework of Erdös-Rényi network (where the probability of having a connection between ith and jth node is p, and, $0 \leq p \leq 1$) and scale free network (where the degree distribution of the network follows a power law) we study the possible emergent dynamics like amplitude death, frequency synchronisation etc in such systems. The subset of m oscillators of lower timescale in the network is defined as S. The equation governing the dynamics of the i_{th} node is given by

$$\dot{X_i} = \tau_i F(X_i) + G\varepsilon\tau_i \sum_{j=1}^{N} A_{ij}(X_j - X_i) \tag{1}$$

where $\tau_i = \tau$ if $i \in$ S, $\tau_i = 1$ otherwise. X_i is n-dimensional and G is an n x n matrix which decides which variables are to be coupled. We take G = diag(1, 0, 0 ..) which means x variable of the i^{th} oscillator is coupled diffusively with the x variable of j^{th} oscillator. ε is the coupling strength which is homogeneous throughout the network. A_{ij} is adjacency matrix of the network.

The results presented here are for dynamics with each node as a Rössler system given by

$$\begin{aligned} \dot{x_i} &= (-y_i - z_i) \\ \dot{y_i} &= (x_i + ay_i) \\ \dot{z_i} &= (b + z_i(x_i - c)) \end{aligned} \tag{2}$$

This intrinsic dynamics is periodic with parameters chosen as a=0.1, b=0.1 and c=4.

2 Results

One of the main results in our work is the phenomenon of amplitude death (AD) that happens under sufficient mismatch in time scales and strong coupling. This corresponds to a state when all the systems go to a synchronized fixed point in time which is identified as the state when the average amplitude $< A >$ calculated over a time becomes zero[5].

Considering Erdös-Rényi network with m nodes chosen randomly, in eqn.1 where $A_{ij} = 1$ with a probability p, we take several realizations of the network and compute the fraction of realizations f that go to a synchronized amplitude death state. We study the scaling in f near this transition as a function of p, for different values of m, and find that there is an optimum value of m where amplitude death occurs at the lowest possible p. We also study this phenomenon by taking three types of probability of connections within the network, p_1 denoting the connectivity between slow to slow systems, p_2 that from slow to fast systems and p_3 that from fast to fast systems. In the bipartite network, where p_2 is non zero and $p_1 = p_3 = 0$ we observe AD. However having a positive value of p_1 and p_3 helps the whole network to go to amplitude death state at a lower value of p_2. By varying the time scale mismatch τ over a range from 0.1 to 1 and ε over 0 to 0.2, we identify the region in the parameter plane (τ, ε) where AD occurs. Outside this region for large τ we observe frequency sychronized state, two frequency state etc.

In the case of scale free network also, with m nodes slow similar results are obtained. However as scale free network has widely differing degrees for nodes, or hubs, we investigate the role of hubs as control nodes that can spread the effects of slowness over the network. For this, after the systems are synchronized we make one of the nodes slow and study how soon the other nodes fall out of synchrony in time. This is characterized by their degrees and shortest paths from the slow node. We discuss this for several starting slow nodes present in the network to quantify the importance of that node in the context of spread of slowness.

We have repeated the study for other dynamical systems like chaotic Rössler, Lorenz system and Landau-Stuart oscillator and found qualitatively similar results.

Summary. Our study shows that, mismatch in time scales of different interacting units can affect the group performance of any complex system leading to decay of overall outcomes and even suppression of all activity. We propose this mismatch in dynamical time scales as another method for achieving AD in interacting systems. In Erdös-Rényi network, we quantify the most sparse configuration to achieve the suppression of dynamics in terms of the probability of connections, whereas in scale free network we quantify the role of nodes to achieve the same. With one control node slow we characterize falling out of synchrony in terms of various measures of the network such as degree, centrality, clustering coefficient etc.

One of the authors (K.G) would like to thank Infosys Foundation Travel Award for funding the trip for this conference.

References

1. D. Das and D.S. Ray, Eur. Phys. J. Special Topics **222**, 785 (2013)
2. L. Kay, Chaos **13**, 1057 (2001)
3. K. D. Williams, W. J. Ingram and J. M. Gregory, Journal of Climate **21**, 5076 (2008)
4. S. De Monte, F. dOvidio and E. Mosekilde, Phys. Rev. Lett. **90**, 054102 (2003)
5. K. Gupta and G.Ambika, Eur. Phys. J. B **89**, 147 (2016)

Parameter Estimation and State Forecasting in Meteorological Models

Inés P. Mariño[1], Sara Pérez-Vieites[2], and Joaquín Míguez[2]

[1] Department of Biology and Geology, Physics and Inorganic Chemistry,
Universidad Rey Juan Carlos, C/ Tulipán s/n, 28933 Móstoles, Madrid, Spain,
ines.perez@urjc.es,
[2] Department of Signal Theory & Communications, Universidad Carlos III de Madrid,
Avenida de la Universidad 30, 28911 Leganés, Madrid, Spain.

1 Introduction

Many problems in the geophysical sciences demand the ability to calibrate the parameters and predict the time evolution of complex dynamical models using sequentially-collected data. Here we introduce a general methodology for the joint estimation of the static parameters and the forecasting of the state variables of nonlinear, and possibly chaotic, dynamical systems. It aims at recursively computing the sequence of joint posterior probability distributions of the unknown model parameters and its state variables conditional on the available observations, possibly incomplete and contaminated by noise.

The new framework combines a Monte Carlo scheme to approximate the posterior distribution of the fixed parameters with filtering (or data assimilation) techniques to track and predict the distribution of the state variables.

2 Model

In order to demonstrate the method, we apply it to a stochastic version of the two-scale Lorenz 96 model. It consists of two sets of dynamic variables that display some key features of atmosphere dynamics [1]: the slow variables $x_j(t)$, $j = 0, \ldots, d_x - 1$ and the fast variables $z_l(t)$, $l = 0, ..., d_x L - 1$ (notice that there are L fast variables per each slow variable). The dynamic variables are assumed to be arranged on a circular structure, hence the operations on the j indices are modulo d_x and operations on the l indices are modulo L. The set of stochastic differential equations (SDEs) can be written as

$$
\begin{aligned}
dx_j &= \left(-x_{j-1}(x_{j-2} - x_{j+1}) - x_j + F - \tfrac{HC}{B} \textstyle\sum_{l=(j-1)L}^{Lj-1} z_l \right) dt + \sigma_x dW_j^x, \\
dz_l &= \left(-CBz_{l+1}(z_{l+2} - z_{l-1}) - Cz_l + \tfrac{CF}{B} + \tfrac{HC}{B} x_{\lfloor \frac{l-1}{L} \rfloor} \right) dt + \sigma_z dW_l^z, \qquad (1)
\end{aligned}
$$

where the time dependence of the variables and noise is left implicit, F is a forcing parameter that controls the turbulence of the chaotic flow, C determines the time scale of the fast variables $\{z_l\}_{l \geq 0}$, H controls the strength of the coupling between the fast and slow variables and B determines the amplitude of the fast variables. W_j^x and W_l^z are independent standard Wiener processes scaled by the constant non-negative factors σ_z and σ_z, respectively. Observations can only be collected from this system once every T time units and only one out of K slow variables can be observed. Therefore, the observation process has the form

$$y_n = \begin{bmatrix} x_{K,nT} \\ x_{2K,nT} \\ \vdots \\ x_{d_yK,nT} \end{bmatrix} + v_n, \tag{2}$$

where $n = 1, 2, ...$ and v_n is a sequence of i.i.d. random variables with common Gaussian density $\mathcal{N}(v_n|0, \sigma_y^2 \mathbf{I}_{d_y})$. The probability density function of y_n given the state $x_n = [x_{1,n}, ..., x_{n,d_x}]$ is denoted $g_n(y_n|x_n)$.

In the computer experiments, the described system is often employed to generate both ground-truth values for the slow variables and synthetic observations. As a forecast model for the slow variables it is common to use the difference equation:

$$dx_j = \big(-x_{j-1}(x_{j-2} - x_{j+1}) - x_j + F - \ell(x_j, \mathsf{a})\big) + \sigma_x dW_j^x, \quad j = 0, ..., d_x - 1, \tag{3}$$

where $\ell(x_j, \mathsf{a})$ is an ansatz for the coupling term. We choose a second order polynomial in x_j, characterized by the roots a_1 and a_2. We discretize the model (3) using a Runge-Kutta method of 4th order (RK4) in order to apply the proposed algorithm. This yields a sequence of states $x_{j,n} = x_j(t = nh)$ where h is the integration step-size and t denotes continuous time.

3 Algorithm

To estimate the unknown parameters we model them as random variables and then use the nested filtering (NF) scheme proposed in [2] to approximate their posterior probability distribution (PD). Let $\theta = [F, \mathsf{a}_1, \mathsf{a}_2]$ and let $\mu_n(d\theta)$ be the PD of θ given the observations $y_1, \ldots, y_n$. Similarly, let $\pi_{n,\theta}(dx_n)$ be the PD of the states x_n conditional on $y_1, \ldots, y_n$, and let $\xi_{n,\theta}(dx_n)$ be the PD of x_n conditional on $y_1, ..., y_{n-1}$. We recursively compute estimates of μ_n and $\pi_{n,\theta}$, $t = 1, 2, ...$, using the general scheme below.

1. Initialization: Draw $\theta_0^{(i)}, i = 1, \ldots, N$, i.i.d. samples from $\mu_0(d\theta)$.
2. Recursive step
   (a) For $i = 1, \ldots, N$:
      i. Draw $\bar{\theta}_n^{(i)}$ from a Markov kernel $\kappa_N(d\theta|\theta_{n-1}^i)$.
      ii. Use an arbitrary filter to approximate $\hat{\xi}_{n,\bar{\theta}_n^{(i)}}$. In the computer experiments, we use a simple extended Kalman filter.
      iii. Compute the normalized weight $w_n^{(i)} \propto \int g_n(y_n|x_n)\xi_{n,\bar{\theta}_n^{(i)}}(dx_n)$.
   (b) Resample to obtain the approximation $\mu_n^N(\theta) = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_n^{(i)}}(d\theta)$.

4 Numerical results

We have conducted computer simulations with integration step $h = 10^{-3}$ and model parameters $F = 8$, $H = 0.75$, $C = 10$ and $B = 15$. We assume that there are $L = 10$ fast variables per slow variable, hence the total dimension of the model is $10d_x$. The noise scaling factors are $\sigma_x = \frac{h}{4} = 0.25 \times 10^{-3}$ and $\sigma_o = 4$, both assumed known. We assume that half of the slow variables are observed in Gaussian noise, i.e., $K = 2$.

Figure 1 shows the true state trajectories, together with their estimates, for the first two state variables of the two-scale Lorenz 96 model. We note that the first variable, x_1, is observed in Gaussian noise, while the second variable, $x_2(t)$, is not observed.

Figure 2 displays the estimates of the fixed parameters F, a_1 and a_2 in the forecast model, together with their reference values.

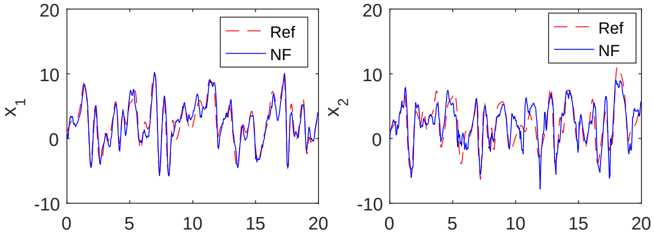


Fig. 1. State values (dashed red line) and their estimates (blue line) for x_1 and x_2.

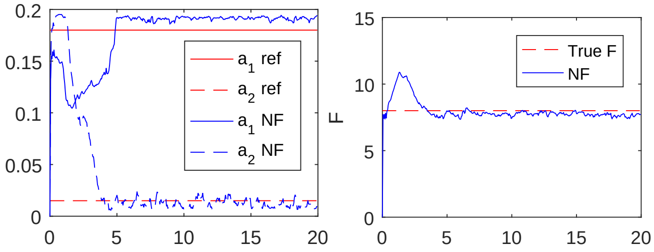


Fig. 2. Estimates of the parameters $\mathsf{a} = [\mathsf{a}_1, \mathsf{a}_2]^\top$ and F in a 4,000-dimensional Lorenz 96 model. The reference values are represented in red dashed lines.

References

1. Arnold, H.M.: Stochastic parametrisation and model uncertainty. Ph.D. thesis, University of Oxford (2013)
2. Pérez-Vieites, S., Mariño, I.P., Míguez, J.: A probabilistic scheme for joint parameter estimation and state prediction in complex dynamical systems. arXiv:1708.03730 (2017)

Network effects on coordination in asymmetric games

Joris Broere[1], Vincent Buskens[1] Jeroen Weesie[1], and Henk Stoof[2]

[1] Utrecht University, Department of Sociology/ICS, Utrecht, the Netherlands
j.j.broere@uu.nl,
[2] Utrecht University, Institute for Theoretical Physics, Utrecht, the Netherlands

1 Introduction

In this paper we study the influence of network structure on global and local behavior in iterated asymmetric 'Battle of the sexes' (BoS) coordination games using myopic best response dynamics. Many computational studies have shown that the spatial structure of a network can have an influence on the evolution of behavior in different types of games [1] [2] [3] [4] [5]. Some research has been done on Battle of the Sexes types of games in the context of spatially distributed interactions in both theoretical and experimental settings, mostly in the context of homogeneous spatial structures such as cellular automaton [6] [7] [8] [9]. To the best of our knowledge, no studies have been performed on the influence of the spatial structure of a network on the equilibrium behavior in a BoS game. We believe that studying a BoS game is particularly interesting because these games provide us with information on which types of nodes end up in their preferred equilibrium and which do not, dependent on the spatial position. If network structure is of any influence, some nodes should have more powerful positions in the sense that their position in the network more easily coordinates on the preferred behavior.

We perform a computational study in which actors play 2×2 games against their neighbors represented by the nodes and edges of a network. In Table 1 the utility matrix of the 2×2 BoS game considered in this study is presented. The row player has the highest payoff when both players choose α and the column player has the highest payoff when both players choose β. Coordinating on the same behavior is more rewarding than miscoordination. In the network version each node is assigned an identity, row player (α prefrence) and column players (β prefrence). The preferences of the nodes are randomly assigned. The constraint of half α prefrence and half β prefrence players is imposed to maximize the coordination problem. On each network we sample 100 different initial conditions, that is different distributions of α and β prefrence players. Assuming that the behavior of most influential or powerful network positions are more likely to converge to their preferred equilibrium than the less influential or powerful network positions, spatial effects can be understood by the probability of a given node

Table 1. Payoff table in BoS, where $0 < S < 1$

	α	β
α	1,S	0,0
β	0,0	S,1

in a network to end up in its preferred equilibrium, irrespective of other initial conditions. So, if different distributions of preferences on a network are played, what is the proportion of times a node converges to the preferred equilibrium given its position in the network.

Three types of networks are considered, namely random Erdös-Rényi (ER) networks, small-world (SW) networks with different rewiring probabilities, and preferential attachment (PA) networks. Several node and network characteristics are considered to be indicative of the behavior, such as global modularity and local centrality measures. We choose networks of size 20 in this study because with this size the relative influence of one node on the global behavior will be substantial. We will also show that most results found are generalizable to larger network sizes.

2 Results

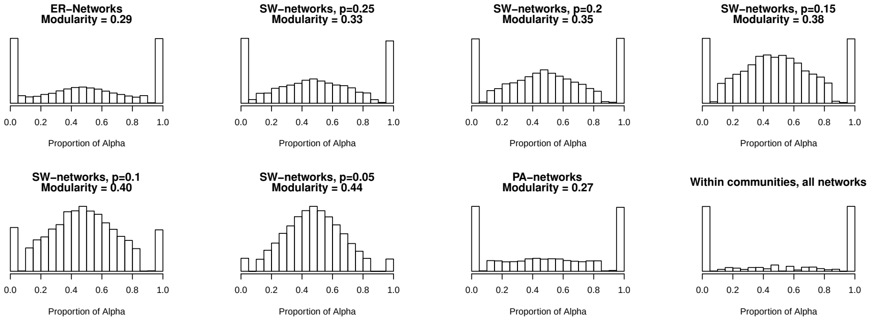


Fig. 1. Proportion of α played in a network after convergence for ER-networks, SW-networks with rewiring probability 0.25, SW-networks with rewiring probability 0.2, SW-networks with rewiring probability 0.15, SW-networks with rewiring probability 0.1, SW-networks with rewiring probability 0.05, PA-networks and within communities of all networks.

As can be seen from Figure 1, random (Erdös-Rényi) networks mostly converge to homogeneous behavior, but the higher the modularity in the network the more heterogeneous the behavior becomes. The Pearson correlation between the heterogeneity of behavior in the network and the modularity of the network is $r = 0.50$. Behavior within the communities of the network is almost exclusively homogeneous. The findings suggest that clustering of networks facilitates self-organization of uniform behavior within clusters, but heterogeneous behavior between clusters.

As can be seen from Table 2, at the local level we find that some nodes are more important in determining the equilibrium behavior than other nodes. Degree centrality is for most networks the main predictor for the behavior and nodes with an even degree have an advantage over nodes with an uneven degree in dictating the behavior. We conclude that the behavior is difficult to predict (Erdös-Rényi) networks and that the network imposes the behavior as a function of clustering and degree heterogeneity in other networks.

Table 2. Regression results, standardized, dependent variable *Power*.

	ER	SW $p = 0.25$	SW $p = 0.20$	SW $p = 0.15$	SW $p = 0.1$	SW $p = 0.05$	PA
Even	0.154	0.159	0.160	0.165	0.165	0.172	0.155
DegC		0.210	0.197	0.191	0.166	0.155	0.315
EVC	0.025						
BetC	0.964						
ClosC	−0.470						
BetC:EVC	−0.839						
EVC:ClosC	0.369						
Constant	0.627	0.478	0.488	0.499	0.516	0.524	0.452
N	20,000	4,000	4,000	4,000	4,000	4,000	20,000
R^2	0.628	0.726	0.714	0.739	0.740	0.757	0.532

**Even = variable indicating an even degree, EVC= Eigenvector centrality, BetC= Betweenness centrality, DegC = Degree centrality, ClosC= Closeness centrality. *Interaction in uncentered variables.*

Summary. All results together seem to indicate that in ER-networks the global behavior is mostly homogeneous, but difficult to predict because often random network characteristics determine to which behavior the network converges. In PA-networks the behavior is also mostly homogeneous, however in this type of network the behavior is dictated by a few influential nodes with high degree centrality. In SW-networks with high clustering, degree centrality is also important, but the spread of behavior is limited by the a node's community, leading to heterogeneous global behavior. In all networks, nodes with an even degree have an advantage over nodes with an uneven degree, because it is easier for nodes with an even degree to obtain a local majority.

References

1. Szabo, G. & Fath, G. Evolutionary games on graphs. Phys. Reports 446, 97216 (2007).
2. Roca, C. P., Cuesta, J. A. & Sanchez, A. Effect of spatial structure on the evolution of cooperation. Phys. Rev. E 80, 046106 (2009).
3. Watts, D. J. & Dodds, P. S. Influentials, networks, and public opinion formation. J. Consumer Res. 34, 441458 (2007).
4. Easley, D. & Kleinberg, J. Cascading behavior in networks. Networks, Crowds, Mark. Reason. About a Highly Connect. World. Camb. Univ. Press. (2010).
5. Santos, F. C., Pacheco, J. M. & Lenaerts, T. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. Proc. Natl. Acad. Sci. United States Am. 103, 34903494 (2006).
6. Alonso-Sanz, R. Self-organization in the battle of the sexes. Int. J. Mod. Phys. C 22, 111 (2011).
7. Hernandez, P., Martnez-Canovas, G., Munoz-Herrera, M. & Sanchez, A. Equilibrium characterization of networks under conflicting preferences. Econ. Lett. 155, 154156 (2017).
8. Hernandez, P., Munoz-Herrera, M. & Sanchez, A . Heterogeneous network games: Conflicting preferences. Games Econ. Behav. 79, 5666 (2013).
9. Mas, M. & Nax, H. H. A behavioral study of noise in coordination games. J. Econ. Theory 162, 195208 (2016).

Online Centrality in Temporally Evolving Networks

Ferenc Béres András A. Benczúr

Institute for Computer Science and Control of the
Hungarian Academy of Sciences (MTA SZTAKI),
{beres,benczur}@sztaki.hu

1 Introduction

A wide range of centrality measures can identify relevant nodes in a graph, however most of them cannot handle networks with rapid dynamics. Conventional methods like PageRank [4], Degree or Katz-index [3] cannot differentiate between links related to recent or past information propagation events. Recently, Rozenshtein et al. [6] proposed an online updateable, dynamic graph centrality measure based on time-respective walks. We propose a method that is similar in kind, but incorporates arbitrary decay functions to model how the impact of information decays over time.

2 Online Centrality: our method

Both Katz-index [3] and PageRank [4] can be considered as path counting centrality measures. Our goal is to define a path count that is updateable by the edge stream of a dynamic network and that can incorporate the actual elapsed time as decay in the propagation function.

For each node v, we maintain the **Online Centrality score** $r(v)$, initialized as a constant α for the initial node set. For each edge uv, we maintain the weight $w(uv)$ and the last activation time $t(uv)$, initially all 0 and $-\infty$, respectively. Next, as edges appear one by one, we update r, w and t as follows. If edge uv appears at time T, first we update $r(u)$ as

$$r(u) := \alpha + \sum_{zu \in A(T)} w(zu) \cdot \varphi(T - t(zu)) \tag{1}$$

where $A(T)$ is the set of edges created before time T and φ is a time decay function that vanishes in infinity. Next we set $w(uv) := r(u)$ and $t(uv) := T$ to propagate the centrality score along edge uv. Note the lazy evaluation: for a given node u, equation (1) is executed only when $r(u)$ needs to be accessed, however all $w(zu)$ are set at the time of change at the other nodes z.

We use Exponential and Rayleigh [5] time decay functions where the intensity of the decay is controlled by a normalization parameter n, see Table 1.

Model	Formula
Exponential (*Exp*)	$\varphi(x) := b^{\frac{x}{n}}$ $(0 < b < 1)$
Rayleigh (*Ray*)	$\varphi(x) := \frac{1}{\sigma^2} \cdot \frac{x}{n} \cdot e^{-\frac{1}{2\sigma^2} \cdot \left(\frac{x}{n}\right)^2}$

Table 1. Time decay formulas incorporated as Online Centrality weight functions φ.

The 6^{th} International Conference on Complex Networks & Their Applications. Nov. 29 - Dec. 01, 2017, Lyon (France)

3 Data set

We collected tweets for the Roland-Garros 2017 between May 23 and June 16, which cover all events[1] of *Day 1–15*. Our data set consists of $444,328$ tweets, containing $351,692$ mentions with their timestamps. By representing these mentions as directed edges between Twitter accounts, we get a dynamic mention network. An edge (u,v,t) corresponds to the event when user u mentions v in his posted message at time t.

We semi-automatically selected the Twitter accounts of the tennis players who participated in the recent French Open tennis tournament. Our method is based on calculating edit distance between the name of professionals in the official schedule and the Twitter account names or the displayed profile names. This way, we managed to assign Twitter accounts to 351 out of 483 professionals who played in Men's, Women's Single/Double, or Legends below 45 categories during the contest. We use these accounts to test our model in the task of daily tennis player prediction.

4 Results

In this work, we examined the problem of central node prediction in the temporally evolving mention network of our Roland-Garros 2017 collection. We considered Twitter accounts of tennis players who participated in rounds on the given day as relevant. For evaluation we used hourly and daily snapshots of Days 1–15, with the daily snapshots ending at UTC+2 midnight. All timestamps are UTC+2, the timezone of the event venue, Paris, France. At the end of each (daily or hourly) snapshot, we update the score vector r for all users based on the given timestamp and store for evaluation.

Our evaluation metric corresponds to identifying active tennis players for a given day, as early as possible. In our experiments, we use NDCG [1] with the relevance function that is 1 if the player participated in tournaments of the given time period and 0 otherwise. In order to show the potential of our model, we used a recently proposed dynamic centrality measure (Temporal PageRank [6]) and various static baselines (PageRank, indegree, Harmonic Centrality [2], negative β-measure [2]) for comparison. In our measurements we use time windows of various sizes to enable static models to track network dynamics. Followers are not available for this data set, hence models designed for follower network analysis such as [7] cannot be applied.

In Fig. 1, we present the daily average NDCG for each centrality measure, using optimal parameter setting, in every hour before UTC+2 midnight of the given day. The results show that for both time decay functions our model (Rayleigh with $n = 2$ hours and Exponential with $n = 1$ hour) significantly outperforms all baseline method from snapshot -23 to -11. Hence in the early stage of the daily information flow we are able to give more accurate predictions about daily tennis players with Online Centrality than with snapshot based measures or Temporal PageRank. Our code base and the data set is available on GitHub[2].

[1] For the full program of Roland-Garros 2017, visit `http://www.rolandgarros.com/en_FR/scores/schedule/`

[2] All our codes as well as the mention graph with the ground truth labels are publicly available here `http://github.com/ferencberes/online-centrality`

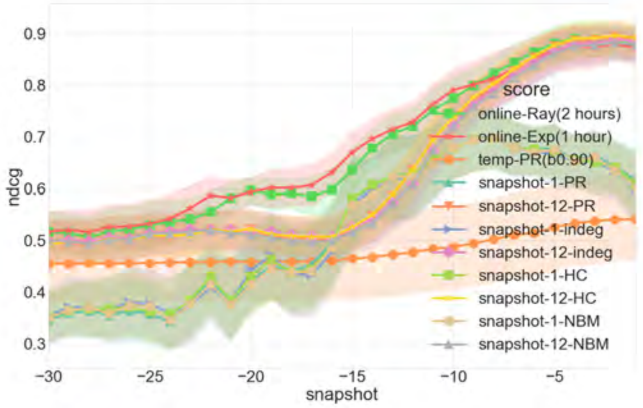


Fig. 1. Daily average NDCG of models with optimal parameters for every hour before UTC+2 midnight (e.g. snapshot -1 corresponds to 23:00 UTC+2). From snapshot -23 to -11 Online Centrality (Online-Ray, Online-Exp) significantly outperforms the baselines. The notations for baseline models are as follows: temp-PR: Temporal PageRank, PR: PageRank, indeg: indegree, HC: Harmonic Centrality, NBM: negative β-measure.

Summary: *We compared our Online Centrality model with snapshot based measures as well as with Temporal PageRank which is, to the best of our knowledge, the only known dynamically updateable centrality algorithm. We measured how accurately and quickly can these models extract emerging new important nodes of the underlying Twitter mention network of Roland-Garros 2017. In our measurements we considered the accounts of tennis players who participated in rounds on the given day as relevant. Our method outperformed the baselines.*

References

1. Al-Maskari, A., Sanderson, M., Clough, P.: The relationship between IR effectiveness measures and user satisfaction. In: Proc. SIGIR. pp. 773–774. ACM (2007)
2. Boldi, P., Vigna, S.: Axioms for centrality. Internet Mathematics 10(3-4), 222–262 (2014)
3. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (1953)
4. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Tech. Rep. 1999-66, Stanford University (1998)
5. Rodriguez, M.G., Leskovec, J., Balduzzi, D., Schölkopf, B.: Uncovering the structure and temporal dynamics of information propagation. Network Science 2(1), 26–65 (2014)
6. Rozenshtein, P., Gionis, A.: Temporal pagerank. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 674–689. Springer (2016)
7. Saez-Trumper, D., Comarela, G., Almeida, V., Baeza-Yates, R., Benevenuto, F.: Finding trendsetters in information networks. In: Proc. SIGKDD. pp. 1014–1022. ACM (2012)

Topology Reconstruction of Dynamical Networks via Constrained Lyapunov Equations

Henk J. van Waarde, Pietro Tesi, and M. Kanat Camlibel

The authors are with the Faculty of Science and Engineering, University of Groningen, The Netherlands.
h.j.van.waarde@rug.nl, p.tesi@rug.nl, m.k.camlibel@rug.nl.

1 Introduction

During the last decades, there has been an increasing interest in networks of dynamical systems. The interconnection structure of such networks is often represented by a graph, where each vertex (or node) of the graph corresponds with a dynamical system, and edges represent interaction between systems. Networks of dynamical systems appear in many contexts, including power networks, water distribution networks, and biological networks.

It is well-known that the overall behavior of a dynamical network is greatly influenced by its network structure (also called network topology). Based on the network structure, predictions can be made on the (steady-state) behavior of a dynamical network. Furthermore, using the network structure, *controllability properties* of the network can be deduced, and control strategies can be determined.

Unfortunately, the interconnection structure of dynamical networks is often not directly available. For instance, in the case of wireless sensor networks the locations of sensors, and hence, communication links between sensors is not always known. Other examples of dynamical networks with unknown network topologies are encountered in biology, for instance in neural networks [8].

Consequently, the problem of *network reconstruction* is studied in the literature (see, e.g., [4], [6], [9], and the references therein). The aim of network reconstruction (also called topology identification) is to find the network structure of a dynamical network, using measurements obtained from the network.

In this abstract, we consider network reconstruction for *deterministic* networks of linear dynamical systems. In contrast to papers studying network reconstruction for specific network dynamics such as consensus dynamics [2] and adjacency dynamics [3], we consider network reconstruction for *general* linear network dynamics. It is our aim to infer the unknown network topology of such dynamical networks, from *state measurements* obtained from the network.

2 Problem formulation

Consider an undirected graph $G = (V, E)$, where $V = \{1, 2, \dots, n\}$ is the vertex set, and E is the set of edges. Associated with G, we consider the network system

$$\begin{aligned} \dot{x}(t) &= Xx(t) \text{ for } t \in \mathbb{R}_{\geq 0} \\ x(0) &= x_0, \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the *state* of the network, and the interconnection matrix X is contained in the so-called *qualitative class* $\mathscr{Q}(G)$ of matrices carrying the structure of the graph G. Examples of members of the qualitative class are the *Laplacian* and *adjacency* matrices associated with the graph G. We denote the state trajectory of (1) by $x_{x_0}(\cdot)$. In this work, we assume that the interconnection matrix X (and graph G) is *unknown*, but the state of the network $x_{x_0}(t)$ can be *measured* for $t \in [0,T]$, where $T \in \mathbb{R}_{>0}$. It is our goal to reconstruct X (and hence, G) using these measurements. Of course, it is possible that multiple *different* network systems of the form (1) generate the *same* measurements $x_{x_0}(t)$ for $t \in [0,T]$. If, on the other hand, the measurements $x_{x_0}(t)$ for $t \in [0,T]$ correspond to a *unique* network system (1), we say the network reconstruction problem is *solvable* for $(x_0, X, \mathscr{Q})$ (for a more formal definition we refer to Definition 1 of [9]).

The first problem of this abstract is to find necessary and sufficient conditions under which the network reconstruction problem is solvable for $(x_0, X, \mathscr{Q})$. Secondly, we want to find a method to reconstruct the matrix X using the measurements $x_{x_0}(t)$ for $t \in [0,T]$.

3 Results and discussion

In this section, we discuss our main results. Firstly, we state necessary and sufficient conditions under which the network reconstruction problem is solvable.

Theorem 1 (Theorem 4 and Proposition 6 of [9]). *The network reconstruction problem is solvable for $(x_0, X, \mathscr{Q})$ if and only if the matrix*

$$P := \int_0^T x_{x_0}(t) x_{x_0}(t)^\top dt \tag{2}$$

is nonsingular.

The matrix P can be obtained from the measurements $x_{x_0}(t)$ for $t \in [0,T]$. Hence, we can check whether the measurements correspond to a *unique* network system (1) by computing the rank of P. In addition to Theorem 1, we state the following theorem, which provides a method to reconstruct the matrix X.

Theorem 2 (Corollary 8 of [9]). *The network reconstruction problem is solvable for $(x_0, X, \mathscr{Q})$ if and only if the Lyapunov equation*

$$SP + PS = Q \tag{3}$$

admits a unique solution S. Under this condition, we have $S = X$.

In other words, if the network reconstruction problem is solvable, we can find the interconnection matrix X (and hence, the graph G) by solving the Lyapunov equation (3). A classical method to solve Lyapunov equations is the *Bartels-Stewart* algorithm [1]. More recently, effort has been made to develop numerical methods for large-scale Lyapunov equations [7].

Theorem 1 and 2 give solutions to the two problems introduced in Section 2. In what follows, we would like to discuss an extension of these results. Suppose that we have additional information on the *type* of network system (1), in the sense that $X \in \mathscr{K}(G)$, where $\mathscr{K}(G) \subset \mathscr{Q}(G)$. That is, suppose that we know that the interconnection matrix X is contained in some (known) subset $\mathscr{K}(G)$ of the qualitative class. Examples of possible sets $\mathscr{K}(G)$ are the sets of Laplacian or adjacency matrices. As before, we are interested in conditions under which the network reconstruction problem is solvable for $(x_0, X, \mathscr{K})$, where in this case we use the symbol $\mathscr{K}$ to indicate that $X \in \mathscr{K}(G)$ for some graph G. It turns out that the network reconstruction problem is solvable for $(x_0, X, \mathscr{K})$ if and only if there exists a unique matrix in the set $\bigcup_{G \in \mathscr{G}_n} \mathscr{K}(G)$ satisfying the Lyapunov equation (3), where the union is taken over the set of graphs $\mathscr{G}_n$ of n nodes (see Theorem 5 of [9]). This means that, in general, for the solvability of the network reconstruction problem for $(x_0, X, \mathscr{K})$, the solution to the Lyapunov equation (3) does not have to be unique (as long as there is a unique solution in the set $\bigcup_{G \in \mathscr{G}_n} \mathscr{K}(G)$).

An attractive feature of the above procedure is that the unique solution is obtained from the Lyapunov equation *if and only if* the network reconstruction problem is solvable (and no additional assumptions, like network sparsity [5], have to be made). Moreover, Lyapunov equations can be solved very efficiently using techniques from [7]. Nonetheless, the approach has some limitations. Specifically, system (1) does not take into account possible noise in the system dynamics and in the measurements. Future work should hence focus on extending the results to stochastic dynamical systems, possibly by using the covariance matrix of the system instead of the matrix P in (2). Moreover, it is of interest to consider nonlinear dynamics instead of the linear dynamics (1).

References

1. R. H. Bartels and G. W. Stewart. Solution of the matrix equation AX + XB = C. *Communications of the ACM*, 15(9):820–826, 1972.
2. G. Chowdhary, M. Egerstedt, and E. N. Johnson. Network discovery: An estimation based approach. In *Proceedings of the American Control Conference*, San Francisco, California, USA, 1076–1081, 2011.
3. M. Fazlyab and V. M. Preciado. Robust topology identification and control of LTI networks. In *Proceedings of the IEEE Global Conference on Signal and Information Processing*, Atlanta, Georgia, USA, 918–922, 2014.
4. D. Materassi and M. V. Salapaka. On the problem of reconstructing an unknown topology via locality properties of the Wiener filter. *IEEE Transactions on Automatic Control*, 57(7):1765–1777, 2012.
5. B. M. Sanandaji, T. L. Vincent, and M. B. Wakin. Exact topology identification of large-scale interconnected dynamical systems from compressive observations. In *Proceedings of the American Control Conference*, San Francisco, California, USA, 649–656, 2011.
6. S. Shahrampour and V. M. Preciado. Topology identification of directed dynamical networks via power spectral analysis. *IEEE Transactions on Automatic Control*, 60(8):2260–2265, 2015.
7. V. Simoncini. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM Journal on Scientific Computing*, 29(3):1268–1288, 2007.
8. P. A. Valdés-Sosa, J. M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:969–981, 2005.
9. H. J. van Waarde, P. Tesi, and M. K. Camlibel. Topology reconstruction of dynamical networks via constrained Lyapunov equations. https://arxiv.org/pdf/1706.09709.pdf, 2017.

High Clustering Protects Against Catastrophic Collapse

Gwen Spencer

Smith College, Northampton , MA, 01063, USA, gspencer@smith.edu

Keywords: cooperation, clustering, threshold models, social influence, contagion, repeated game-play. **Full paper accepted at "Network Science" (Cambridge Un. Press).*

Introduction. In this computational study, we consider the spread of a binary behavior through a network of *conditional cooperators* (or *moody conditional cooperators*). We propose a novel perspective on a source of advantage that can be accessed by a highly-clustered society. Many previous studies have concentrated on whether network structure can facilitate *outbreaks of cooperation* (extending the pre-network interest in the *initial viability* of a cooperative strategies in a population of selfish players). In contrast, we focus on whether particular network structure can protect against *catastrophic collapse of cooperation* in the presence of *shocks of defection* that impact an initially-widely-cooperative society.

Our systematic suite of experiments tests how the ability to avoid such collapse can be degraded as local links are reallocated to long ties. In particular, we consider a sequence of randomly-realized graphs that result from applying the Watts and Strogatz [6] random-rewiring procedure to a set of small dense communities. This roughly replicates a sequence of stochastic block models in which the clustering coefficient is smoothly eroded from 1 to a small value associated with a graph that approaches a random graph with minimal community structure.

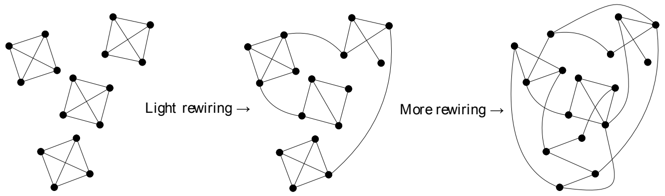


Fig. 1. Schematic of Random Rewiring of Dense Local Communities (Complete Graphs). Clustering coefficient decreases gradually as the probability of rewiring, *p*, increases.

Assuming a time-indexed binary [0,1]-behavior (defection vs. cooperation) and that players apply a threshold-based decision rule to update their behavior, we measure how large a randomly-distributed shock of defection must be to push the network into a state of *catastrophic collapse of cooperation*. We conduct tests for both the classical threshold-based model (that is, each player will cooperate exactly when at least a *h*-fraction of her neighbors cooperated in the previous time step), and an empirically-observed variant in which each player will condition both on the prior behavior of her neighbors and also on *her own behavior in the previous time step*. Such *moody conditional cooperation* has been empirically documented in repeated networked game play of Prisoner's Dilemma Games and Public Goods Games (e.g. [3], [1], [2], [4]).

We have several motivations for considering such a model. The most practical of these motivations is to help understand why several very-prominent human experiments have failed to detect an impact of network topology in encouraging cooperation in repeated game-play in networks. The authors of behavioral studies like [5] and [1] often frame their results as being quite surprising, and claim that their observations discredit the classical theoretical prediction that network structure (that impedes the perfect-mixing of a population) should encourage the emergence of cooperative behavior. Our study points out how several apparent "refutations" of this type may actually be understood as simple consequences of the region of the parameter space tested by particular experimental designs. That is, the findings of these human experiments may be perfectly compatible with topology playing an important role in promoting cooperation in other parts of the parameter space.

Results. We observe a remarkably-linear *protective effect of clustering* coefficient that becomes active above a *critical level of clustering*. Notably, both the critical level and the slope of this dependence is higher for decision-rule parameterizations that correspond to higher *costs of cooperation*. For traditional threshold-based behavior updating, this linear protective effect is quite pronounced above thresholds of 0.5, and becomes increasingly mild as player thresholds for cooperation decreases towards 0.

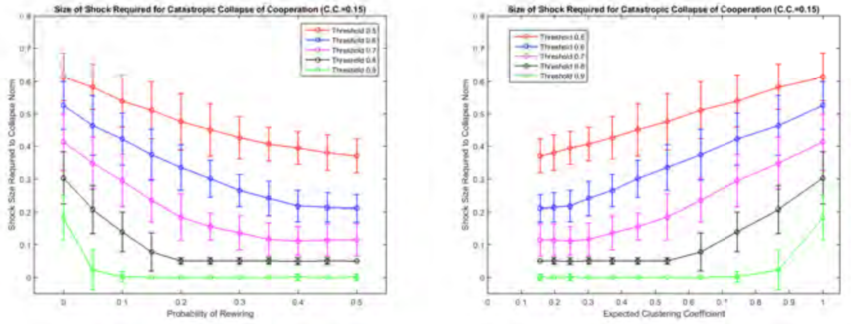


Fig. 2. High Clustering Increases Ability to Withstand Defection Shocks, But Large Shocks Will Cause Catastrophic Collapse Regardless of Topology. *Ability to Withstand Defection Shocks vs. Rewiring Probability (left panel) and vs. Clustering Coefficient (right panel). Five thresholds for cooperation are studied (0.5, 0.6, 0.7, 0.8, 0.9). Twenty initial communities of ten individuals each (total society of 200 nodes) are rewired.*

Similarly, for *moody cooperation* behavior update model, when a population of *Moderate Players* is augmented with *Stingy Players* (and then players are placed uniformly at random in the network), the protective effect of clustering becomes very distinctive. In contrast, when a population of *Moderate Players* is augmented with *Generous Players*, change in the clustering coefficient doesn't improve the already-high level of defection the network can endure (without falling into cooperation collapse). That is, when many players are willing to cooperate even after observing many defections among their neighbor-sets, no protective effect of clustering is predicted by our models.

Our observations appear consistent for small synthetic networks (even when initial small community sizes are non-uniform), and a sizable real-data example. The real network data set we consider describes Co-board-membership in public-limited companies Norway (1,421 individuals). We observe remarkably similar behavior to our smaller synthetic examples even though the real network has a power-law-like degree distribution, while our synthetic examples have binomial degree distributions.

Our framework provides a novel interpretation of the highly-cited behavioral study of Suri and Watts [5] on the role of network topology in repeated game. Suri and Watts considered repeated play of a public-goods game in small networks with different clustering coefficients. The long-term behavior of all topology treatments appears to be very low levels of cooperation. Though Suri and Watts do detect conditional responses (players respond to observations of the past choices of neighbors), they conclude that network structure does not impact levels of cooperation in networked game play.

Our framework suggests a different interpretation: *Suri and Watts simply tested a portion of the parameter space where no impact of topology was predicted.* Suri and Watts document a high percentage of very-stingy contributions in the first round of their game. Roughly 45% of players respond to the Suri and Watts game by choosing a round-1 strategy that could be viewed by their neighbors as “defection.” Under a random spatial distribution of a population of 45% defectors in the exact 24-node networks tested by Suri and Watts, we find that collapse of cooperation is (by far) the most likely outcome for each of the topologies they test. Once networked games are pushed into a state of *catastrophic collapse of cooperation*, their behavior is extremely hard to distinguish. Further, any variation between topologies is well-obscured by the substantial variation for a fixed topology where the spatial placement of defecting round-1 players is not controlled (we know of no human studies that control initial spatial distribution).

Prominent human studies on *moody conditional cooperation* also make bold claims that topology is irrelevant for cooperation [1]. Again we find that there are simple parameter-value-based explanations about why all topologies tested should converge to catastrophic collapse of cooperation. We suggest that such issues might be remedied by an advance stage of experimental design in which scientists probe the round-1 behavior of players and manipulate the local game reward structure so that some impact of topology is actually predicted computationally for the master experiment.

References

1. Gracia-Lzaro, C., Ferrer, A., Ruiz, G., Tarancn, A., Cuesta, J.A., Snchez, A., Moreno, Y.: Heterogeneous networks do not promote cooperation when humans play a prisoners dilemma. Proceedings of the National Academy of Sciences 109(32), 12922–12926 (2012)
2. Grujić, J., Gracia-Lázaro, C., Milinski, M., Semmann, D., Traulsen, A., Cuesta, J.A., Moreno, Y., Sánchez, A.: A comparative analysis of spatial Prisoner’s Dilemma experiments: Conditional cooperation and payoff irrelevance. Scientific Reports 4, 4615 (Apr 2014)
3. Gruji, J., Fosco, C., Araujo, L., Cuesta, J., Snchez, A.: Social experiments in the mesoscale: Humans playing a spatial prisoner’s dilemma. PLOS ONE 5(11), 1–9 (11 2010)
4. Horita, Y., Takezawa, M., Inukai, K., Kita, T., Masuda, N.: Reinforcement learning accounts for moody conditional cooperation behavior: experimental results. Scientific Reports 7, 39275 (2017)
5. Suri, S., Watts, D.: Cooperation and contagion in web-based, networked public goods experiments. PLoS ONE 6(3), e16836 (2011)
6. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature 393(6684), 409–10 (1998)

Dynamical processes modelled by *k*-path Laplacian operators

Ernesto Estrada[1], Ehsan Hameed[1], Naomichi Hatano[2], and Matthias Langer[1]

[1] Department of Mathematics and Statistics, University of Strathclyde,
26 Richmond Street, Glasgow G1 1XH, UK
[2] Institute of Industrial Science, University of Tokyo,
4-6-1 Komaba, Meguro, Tokyo 153-8505, Japan

In this work we focus on dynamical process on networks in which nodes not only interact to its nearest neighbours but also through some long-range influences (LRIs) [1]. Here we study a diffusive process controlled by the generalised k-path Laplacian operators (LO) L_k. The introduction of the k-path LOs can help in conducting more precise studies of a network's dynamics in different applications. The main goal of this research is to study a generalised diffusion equation using the transformed generalised k-path LOs for locally finite infinite networks. First, we proved a few properties of these operators, such as their boundeness and self-adjoitness [2]. We studied three different transformations of the k-path LOs, namely, Laplace, factorial and Mellin. We proved that all three transformed k-path LOs are also in general bounded and seld-adjoint. Finally, we used the transformed k-path LOs to obtain a generalised diffusion process for an infinite path graph. We prove analytically here that under the Mellin transform of these k-path LOs for certain values of the parameter, $1 < s < 3$, a superdiffusive dynamics appears (see Fig. 1). On the contrary, the generalized diffusion equation using Laplace and Factorial transformed operators always produce normal diffusive processes.

References

1. E. Estrada: Path Laplacian matrices: introduction and application to the analysis of consensus in networks. *Linear Algebra Appl.* 436 (2012), 3373–3391.
2. E. Estrada, E. Hameed, N. Hatano, M. Langer: Path Laplacian operators and superdiffusive processes on graphs. I. One-dimensional case. *Linear Algebra Appl.* 523 (2017), 307–334.

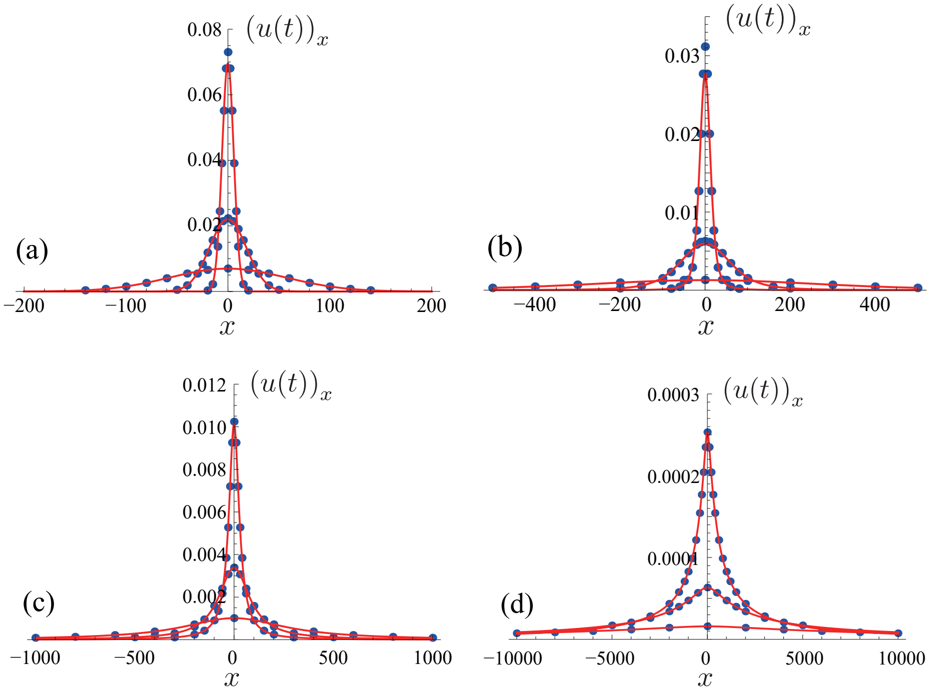


Fig. 1. The time evolution of the density profile under the Mellin-transformed k-path Laplacian: (a) $s = 4$ for $t = 10, 100, 1000$ from high to low; (b) $s = 2.5$ for $t = 10, 100, 1000$ from high to low; (c) $s = 2$ for $t = 10, 30, 100$ from high to low; (d) $s = 1.5$ for $t = 10, 20, 40$ from high to low. In every panel, the blue dots indicate the result of numerical integration, whereas the red curves indicate the asymptote.

Cascading collapse of online social networks

János Török[1,2] and János Kertész[1,2]

[1] Center for Network Science, Central European University,
Nádor u. 9, H-1051 Budapest, Hungary
[2] Department of Theoretical Physics, Budapest University of Technology and Economics,
H-1111 Budapest, Hungary
Email: torok@phy.bme.hu

The role of collective effects in spreading on and growth of social networks has been studied widely[16, 3, 1, 13, 4, 5, 7, 12], yet, much less is known about their role in the process of the shrinkage and collapse of such networks[8, 14, 11]. Here we focus on the following questions: What is the mechanism of the contraction of online social networks and how does the leavers influence the churning behavior of the persons remaining with the service.

In this presentation we study the full lifecycle of a social network site (iWiW) which dominated one country for years (2006-2011) and hosted two thirds of the population with internet access[9, 10]. Its collapse was very fast in spite of the efforts of the owners and it was accompanied by the increase of the local popularity of the competing international alternative. The question remains to be answered whether the fall was caused solely by an external event (popularity of facebook) or local collective action of the users.

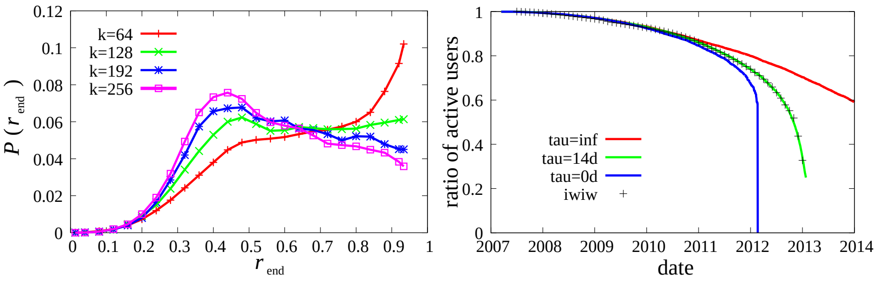


Fig. 1. (a) Distribution of the fraction of active acquaintances at the time of the last login of users for four different user degrees in iWiW. Note that the absence of values less than $r_{\text{end}} \simeq 0.3$ is due to the fact that our database was truncated when 70% of the users left. (b) The cumulative fraction of active users (all users having a last login date after the indicated time) in iWiW (crosses) and in the cascade model (green). For comparison we show also results with zero characteristic leaving time (blue), and with zero threshold, i.e. no social effect (red). Parameters: N=10000, $\langle k \rangle$=10.

We measured r_{end} the fraction of acquaintances that were still active at the time of the last login of a user. If r_{end}=1 the user left before all its friends and if r_{end}=0 then after them. We created histograms for users having different degrees and plotted it in Fig. 1 (a). We see that users with low degree leave mainly early, before their friends while others with large number of friends stay until about 45% of their friends are still active on the site. The transition is sharp and was found to be about $k \simeq 130$.

Our interpretation of this result is that the level of embededness on the site has a crucial impact on the behavior of the users[15]. Those who invested a lot of time in building up their social network on the site stay until most of their friends are active but if more than half of them left they also leave because it is not worth staying any more.

The above idea was put in a model based on threshold mechanism [16, 6]. After creating an underlying social network we consider nodes as users of the site. The external effect is modeled by randomly removing users from the site with probability inversely proportional to their degree. The rate of user removal was found to increase linearly in the system starting from 2007. This linear rate could be obtained by a fit for the early stage of the ratio of active user's curve and was implemented in the model.

Users are given a random individual uncorrelated threshold value in the range $\lambda = 0.5 \pm 0.2$. If the fraction of their active friends drops below this threshold the user leaves the system with a timescale τ. This implements the effect that we do not immediately discover the inactivity of our friends.

The model fits perfectly the numerical data if a network with average degree $\langle k \rangle = 10$ and a user timescale of $\tau = 14$ days are used. This first may sound counter intuitive but it was suggested that the intimate circle of our friends contains $12-15$ people [2] and it seems when considering whether to leave a site we case about only about our intimate friends. We are mainly interested in the news from these people. The 14 days waiting time is a reasonable period to discover the inactivity or our friends.

The model could also reproduce other quantities measured on the data. The most successful was the number of churning users which diverges as a power law if time is counted backwards from the collapse date. A churning user is anyone who left within two weeks after one of its acquaintances. This results enables us to predict in advance the collapse time of a still functioning site. This was repeated for other dataset and an agreement with the real lifecycle was found.

In summary in a world where social influence has higher effect than mainstream media it is of importance to quantify and model it. By studying the collapse of online social network sites we reveal that a user may leave the service for a competing one either due to exogenous or social effects depending on its embededness. The dynamics can be well described by a simple threshold model taking into account these factors. The study of the resulting cascades allows for an early prediction of the collapse time of the site.

References

1. Centola, D.: The spread of behavior in an online social network experiment. Science 329(5996), 1194–1197 (2010)
2. Dunbar, R.I., Arnaboldi, V., Conti, M., Passarella, A.: The structure of online social networks mirrors those in the offline world. Social Networks 43, 39–47 (2015)
3. Gleeson, J.P., Cahalane, D.J.: Seed size strongly affects cascades on random networks. Physical Review E 75(5), 056103 (2007)
4. Gleeson, J.: High-accuracy approximation of binary-state dynamics on networks. Physical Review Letters 107, 068701 (2011)

5. Gleeson, J.: Binary-state dynamics on complex networks: Pair approximation and beyond. Physical Review X 3, 021004 (2013)
6. Karsai, M., Iñiguez, G., Kikas, R., Kaski, K., Kertész, J.: Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading. Scientific Reports 6, 27178 (2016)
7. Karsai, M., Iñiguez, G., Kaski, K., Kertész, J.: Complex contagion process in spreading of online innovation. Journal of The Royal Society Interface 11(101), 20140694 (2014)
8. Kim, H.S., Yoon, C.H.: Determinants of subscriber churn and customer loyalty in the korean mobile telephony market. Telecommunications Policy 28, 751765 (2004)
9. Lengyel, B., Varga, A., Ságvári, B., Jakobi, Á., Kertész, J.: Geographies of an online social network: weak distance decay effect and strong spatial modularity. PLOS ONE (2014)
10. Lengyel, B., Varga, A., Ságvári, B., Jakobi, Á., Kertész, J.: Geographies of an online social network. PloS one 10(9), e0137248 (2015)
11. Mitrović, M., Paltoglou, G., Tadić, B.: Quantitative analysis of bloggers collective behavior powered by emotions. Journal of Statistical Mechanics: Theory and Experiment 2011(02), P02005 (2011)
12. Ruan, Z., Iniguez, G., Karsai, M., Kertész, J.: Kinetics of social contagion. Physical review letters 115(21), 218702 (2015)
13. Singh, P., Sreenivasan, S., Szymanski, B.K., Korniss, G.: Threshold-limited spreading in social networks with multiple initiators. Scientific reports 3 (2013)
14. Tadić, B., Šuvakov, M., Garcia, D., Schweitzer, F.: Agent-based simulations of emotional dialogs in the online social network myspace. In: Cyberemotions, pp. 207–229. Springer (2017)
15. Török, J., Murase, Y., Jo, H.H., Kertész, J., Kaski, K.: What does big data tell? sampling the social network by communication channels. Physical Review E 94, 052319 (2016)
16. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences 99(9), 5766–5771 (2002)

Modeling structure and dynamics of discussion threads in online boards with Hawkes processes

Alexey Medvedev[1,2] and Renaud Lambiotte[1,3]

[1] NaXys, Université de Namur, Namur, B-5000, Belgium
[2] ICTEAM, Université Catholique de Louvain, Louvain-la-Neuve, B-1348, Belgium
[3] Mathematical Institute, University of Oxford, Oxford, UK
an_medvedev@yahoo.com,
WWW home page: http://alexeymedvedev.com

1 Introduction

Online social platforms provide a fruitful source of information about social interaction. Depending on the platform, various tree-like cascading patterns emerge as a consequence of such interaction. For example, on Twitter or on Facebook people interact via resharing messages, which turns into cascade trees of reshares [5], on email networks people forward messages to their peers resulting in trees of email forwards [4], on online boards like Digg or Reddit people interact via discussing particular posts, which leaves a trace of discussion trees. The two main questions arise: what is the shape of these cascades and what is the dynamics of their evolution?

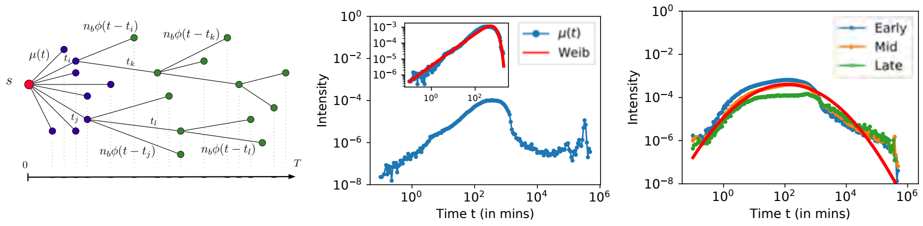
We consider cascades given by discussion trees of posts, which can be viewed as rooted trees, where nodes represent comments and there is the special node, the root, which represents the initial post. Connections represent a 'reply-to' relation between nodes. The content of nodes is disregarded.

The question of evolution of discussion threads is now gradually being understood. In [2, 3] the authors studied only the structural evolution of discussion trees in four large Internet boards, and they suggested a tree generation model based on preferential attachment (PA) mechanism. However the dynamical properties are left out of consideration. In [6] the authors introduce a merely theoretical model which aims to describe structural and temporal evolution of the discussions. Their proposition is to use a specific Levy point process to generate timings, then construct the PA discussion tree assigning to each new node a subsequently generated timing. However, being a sort of a mean-field model, it describes evolution on average, thus having limited utility in practice.

We propose a model of discussion trees generation based on the self-exciting Hawkes processes, which represents both the tree structure and temporal information. We use the dataset of Reddit discussion threads to show that structurally trees resemble Galton-Watson trees with a root bias, and for large trees the dynamics of comments attraction can be well modeled using non-homogeneous Poisson processes.

2 Dataset

The dataset of Reddit discussion threads consists of all posts and comments submitted to Reddit from Jan, 2008 till Jan, 2013. The dataset in total contains more than 120 million posts and around 1.1 billion comments.



(a) (b) (c)

Fig. 1: Schematic structure of a model of discussion tree (a), and aggregated intensity of comment arrival to (b) the root and (c) the comments. The inset of (b) shows the intensity of response times for discussions that last less than 36 hours, depicting the Weibull shape given by the red curve. The intensity for comments on (c) is aggregated into three sets: early comments appeared within 6 hours from post creation, mid created within 6 and 24 hours after the post's creation and late are the rest.

3 Model

The model of discussion trees is based on self-exciting Hawkes processes, which are defined as Poisson point processes with time-dependent intensity

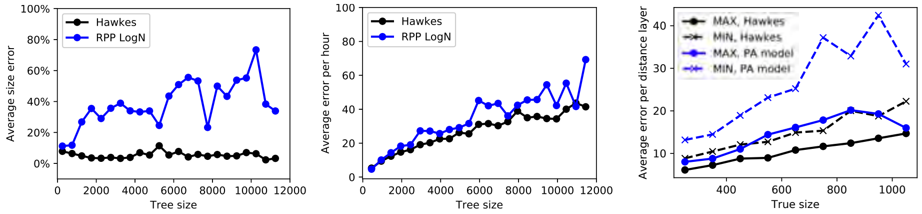
$$\lambda(t) = \mu(t) + n_b \sum_{i:t_i<t} \varphi(t-t_i),$$

where $\mu(t)$ is the background intensity, $\varphi(t)$ is the memory kernel and n_b is the average branching ratio. We assume that $\langle \mu \rangle = \int_0^\infty \mu(t)dt < \infty$ and $\langle \varphi \rangle = 1$, thus the process is almost surely finite if $n_b \leq 1$. The associated process tree has the following straightforward description: let the comments arrive to the root according to the point process with intensity $\mu(t)$, and each comment i, generated at t_i, attracts further comments according to a point process with intensity $n_b\phi(t-t_i)$ (see Figure 1, (a)).

The data shows that functional form of μ resembles the pdf of Weibull distribution $W(a,b,\alpha)(t) = a(\alpha/b)(t/b)^{\alpha-1}\exp(-(t/b)^\alpha)$ and ϕ is the pdf of lognormal distribution $LN(\mu,\sigma)(t) = (1/\sqrt{2\pi}\sigma t)\exp(-(\log(t)-\mu)^2/(2\sigma^2))$ (see Figure 1, (b,c)). For each given tree, the parameters of the distributions are fitted with maximum likelihood estimation (MLE) separately for the root and for the rest of aggregated comments. The parameter n_b is then estimated as the average number of further comments attracted by comments.

4 Baseline models

The structural performance is evaluated in comparison to the *PA model* described in [2]. Temporal aspects of Hawkes model are compared with the Reinforced Poisson *(RP) model* [1], which is shown to outperform linear regression and deterministic point process model in describing and predicting Twitter cascades. Applied to discussion trees, we show that lognormal kernel performs better than the power-law as for Twitter cascades. Parameters of the RP model are also fitted using maximum likelihood estimation.



(a) (b) (c)

Fig. 2: (a) Average size error in percents of true size; (b) average activity error per hour for simulation; (c) average structural error per distance layer.

5 Results

Total size and timings. The model evaluation was performed by considering the top 7000 largest trees from the years 2008, 2010 and 2012. For each tree the parameters were estimated and the statistics of 50 runs of simulations were recorded for both models. Size error is estimated as an absolute relative deviation of the average size of simulated trees from the true value. Error in temporal activity is given as a average absolute difference in total hourly comment arrivals, where the average is taken only over hours, which have comments in one of the considered timelines. As we can see from Figure 2, (b), the mean average activity error per hour grows at a slower rate than the tree size and the errors for both considered models are comparable, the average error of total size estimation is mainly worse for the RP model Figure 2, (a).

Structure. The model evaluation was performed by considering the top 2000 largest trees from the year 2008. The parameters of PA model were prior measured for bunches of trees of evenly distributed sizes, and 50 runs of simulations were performed for both models. Deviations in the tree structure were measured via comparing distance distribution of nodes in both given and simulated trees, and calculating average absolute difference of nodes per layer, where two ways of averaging were considered: only up to the minimum (MIN) or up to maximum (MAX) depth of both trees. The results are shown on Figure 2, (c), where the Hawkes trees receive better scores than the PA ones.

References

1. Gao, S., et.al: Modeling and predicting retweeting dynamics on microblogging platforms. In: Proceedings of WSDM '15. pp. 107–116. WSDM '15 (2015)
2. Gómez, V., Kappen, H.J., Kaltenbrunner, A.: Modeling the structure and evolution of discussion cascades. In: Proceedings of the HT '11. pp. 181–190 (2011)
3. Gómez, V., Kappen, H.J., Litvak, N., Kaltenbrunner, A.: A likelihood-based framework for the analysis of discussion threads. World Wide Web 16(5-6), 645–675 (2013)
4. Iribarren, J.L., Moro, E.: Branching dynamics of viral information spreading. Phys. Rev. E 84, 046116 (2011)
5. Kobayashi, R., Lambiotte, R.: Tideh: Time-dependent hawkes process for predicting retweet dynamics. In: ICWSM' 2016. pp. 191–200 (2016)
6. Wang, C., Ye, M., Huberman, B.A.: From user comments to on-line conversations. In: Proceedings of the KDD '12. pp. 244–252. KDD '12 (2012)

Part VI

Network Models

Synthetic Models for Multi-Layered Networks

Marzena Fügenschuh[1], Ralucca Gera[2], Mitchell Heaton[2], and Tobias Lory[1]

[1] Beuth University of Applied Sciences, Berlin, Germany
fuegenschuh@beuth-hochschule.de
[2] Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA
rgera@nps.edu

1 Introduction

Air transportation networks are known to be of a multiplex structure. These networks develop independently by carrier based on economic and political factors as well as interactions between one another. Creating synthetic models for air transportation networks provides a tool towards a deeper understanding of their structure and development. In particular the multi-layer structure is hard to be imitated. Multilevel or multilayered networks, frequently referred to as multiplexes, have been considered as a detailed extension of the single layered networks [1–3]. This structure is desirable in our case, as each airline company can easily be modeled by a layer, with the airports being captured by the nodes. While generating synthetic networks [4] has been very active research area, less has been done in synthetic multilayered network generation [3]. In the most common approach growing multiplex network models are based on preferential attachment [5–7] as they usually model relations in social networks. To create synthetic models we build on the understanding the structure of air transportation networks [8, 9], in particular, the European Air Transportation Network (EATN). A model for an air transportation network exploiting scale-free structure based on preferential attachment was introduced [10], but it does not exploit the multilayer structure. In [7] the scale-free structure of the network is combined with the multilayer approach and a generative model imitating EATN is presented. It bases on an enhanced preferential attachment method.

We present two models for synthetic generation of multiplex networks following two different approaches.

In the first model, called *StarGen*, we create the multiplex using a modified preferential attachment from [7] to assign edges to different layers. The model focuses on the diversity of the distinct layers within a multiplex. Inspired by *BinBall*'s preferential attachment [7] we create an asynchronous growth of the layers in the multiplex. To do so, we allow different sizes of the layers based on a predefined distribution of layers' edge count, and assign to each layer its own local exponent in the formula for the preferential attachment.

In the second model, we first mimic the structure of each layer separately and then conduct ways to combine the layers to a multiplex. We named this model ANGEL as it stands for Airline Network Generation Emphasizing Layer. Our approach in general is first to distribute the nodes of the multiplex among the layers enforcing their node overlap. Then, each layer grows in the edge size separately, but contributes simultaneously to the whole multiplex due to node sharing across other layers. Our design emphasizes the hub-spoke structure of the European Air Transportation Network (EATN). For that, we generally differentiate between hub and non-hub nodes. In our methodology we first assign non-hub nodes to the layers enforcing overlap of the layers. Then we assign hubs to layers. To achieve overlap on hubs across the layers we create a directed sub-network on all hubs. (If hub u points from layer L to hub v from layer K, then v must belong to L). Finally, we apply the method for layer creation, as a plug-in, to connect hubs and non-hubs within the single layers. Thus we obtain the multiplex.

2 Results

We validate both models with the EATN [**?**] and the *BinBall* model [7] and show two representative statistics capturing the global perception from the multiplexity and the local point of view, the intra-layer structure of the network. In Figure 1 we compare the average degree distribution, the average shortest paths per node, and the average node centrality coefficient over 100 instances of each synthetic multiplex. Both *StarGen* and ANGEL outperform *BinBall*, however ANGEL delivers a slightly better approximation for all values than *StarGen*. Although almost all EATN-layers resemble hub-spoke structure, it shapes differently over the layers. We deduce it from the highly volatile percentage of one degree nodes across the layers, see the first chart on the left in Figure 2. Each color represents a group of nodes: of degree 1, followed by the ones of degree less than t % of local maximum degree, where $t \in \{10, 20, \ldots, 100\}$. For each x-value representing a layer, the y-value is the count of each color group, normalized by the layer's node count. The layers are sorted by the value of the 1-degree group. The remaining

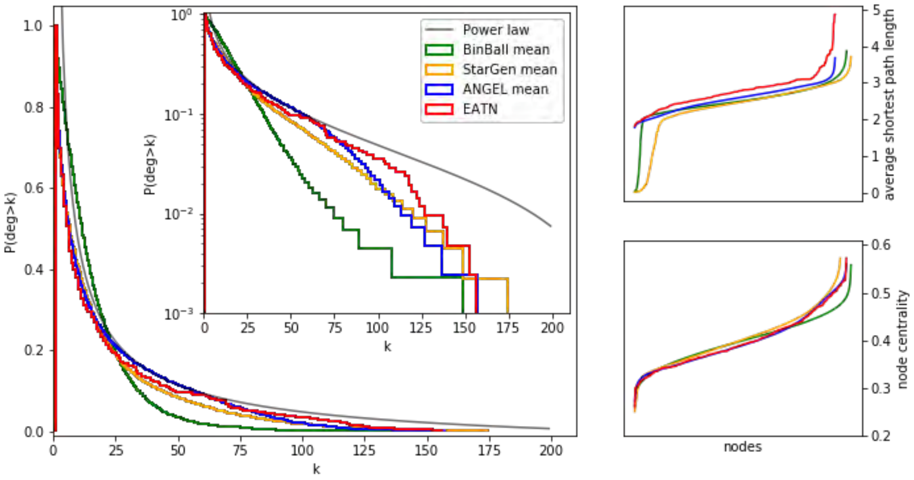


Fig. 1. Left: Degree histogram of the multiplexes with the log scale in the inset. Upper right: average shortest path, lower right: centrality coefficient, per node.

three plots in Figure 2 display degree distribution per layer of one instance of each multiplex model: *BinBall*, *StarGen*, and ANGEL, respectively. Again, the ANGEL-model appears to be the winner.

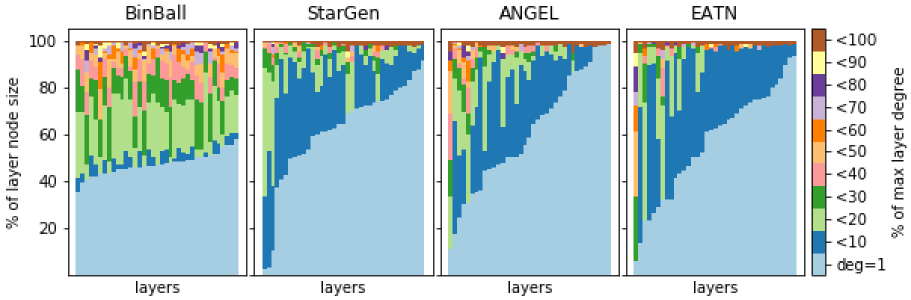


Fig. 2. Color map with layer degree groups: of degree 1, followed by the ones of degree less than $t\,\%$ of layer maximum degree, where $t \in \{10, 20, \dots, 100\}$.

Summary. The *StarGen* algorithm strikes with its simplicity. Using the varied local exponent for each layer we can influence the diversity of the local degree distribution within the multiplex. However, even with experimental fine-tuning we were not able to match EATN as well as ANGEL does. Furthermore, the *StarGen* model neglects the interconnections between the layers. The ANGEL model is more complex, but it benefits from the introduced control of the layer node overlap, and the model can be generalized for arbitrary multilayered networks.

References

1. De Domenico, M., Solé-Ribalta, A., Cozzo, E., and others: Mathematical formulation of multilayer networks. Phys. Rev. X3(4), 2013. doi: 10.1103/PhysRevX.3.041022
2. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., Porter, M. A. : Multilayer networks. J. Compl. Net. 2(3), 2014. doi: 10.1093/comnet/cnu016
3. Boccaletti, S., Bianconi, G., Criado, R., and others: The structure and dynamics of multilayer networks. Phys. Rep. 544(1), 2014. doi: 10.1016/j.physrep.2014.07.001
4. Barthélemy, M. : Spatial networks. Ph. Re. 499(1), 2011. doi: 10.1016/j.physrep.2010.11.002
5. Kim, J. Y.,Goh, K.-I. : Coevolution and correlated multiplexity in multiplex networks. Phys. Rev. Let. 111 (5), 2013. doi: 10.1103/PhysRevLett.111.058702
6. Nicosia, V., Bianconi, G., Latora, V., Barthelemy, M. : Growing multiplex networks. Phys. Rev. Let. 111 (5), 2013. doi: 10.1103/PhysRevLett.111.058701
7. Basu, P., Sundaram, R., Dippel, M. : Multiplex networks: a Generative Model and Algorithmic Complexity. IEEE/ACM 2015. doi:10.1145/2808797.2808900
8. Wilkinson, S.M., Dunn, S., Ma, S. :The vulnerability of the European air traffic network to spatial hazards, Natural Hazards 60(3), 2012. doi:10.1007/s11069-011-9885-6
9. Cardillo, A., Gmez-Gardees, J., Zanin, M., and others: Emergence of network features from multiplexity. Sci. Rep. 3(2) 2013. doi: 10.1038/srep01344
10. Guimer, R., Amaral, L.A.N. : Modeling the world-wide airport network, The European Physical Journal B 38(2), 2004. doi:10.1140/epjb/e2004-00131-0

Complexity and Heterogeneity in dynamical networks and system instability

Fabio Vanni[1], Giovanni Dosi[1], Andrea Roventini[1], and Mauro Napoletano[2]

[1] Sant'Anna School of Advanced Studies, Institute of Economics, Pisa, ITALY
[2] OFCE-Science Po, Nice, France and Universit Cote d 'Azur, SKEMA, CNRS, GREDEQ, Nice, France

1 Introduction

We present an analytical solution for the connectivity of a network formation model with a "non-simultaneous" linking scheme. We show how just a minimal heterogeneity in a network can trigger the arise of non trivial behaviors both in the topological and temporal structure of the system.

In this class of networks introduced by [1] and developed in [3], system evolution has nodes which can either create or destroy links according to their target attitudes. This model belongs to a class of networks not based on the rationale of linking probability between pair of nodes, but it is based on the initiative of the single units when they act. We focus the attention to a minimal heterogeneous population with opposite attitude in the linking formation. From one side there are agents which can be seen as generators of links since their attitude in the networks is to have a virtually infinite degree. From the other side, other nodes can be seen as destroyers since their degree preference is virtually zero. At each time step a node is picked randomly among the network. If a generator is extracted, it acts adding a link with a randomly chosen destroyer with which it has no link yet. If a destroyer is picked, it cuts a link with a randomly chosen generator neighbors. Despite of a such extreme situation, this formation procedure introduces an interesting point of view where the network can be seen as a temporal system which evolves according to a well defined stochastic processes with a non-trivial birth-death structure.

The bipartite structure inserts the most simple situation of heterogeneous agents. According to the heterogeneity level of the network, we can observe and describe the presence of emergent properties in the system consisting in node-degree correlations in the link distribution, and anomalous fluctuations in the time series of the connectivity variable. Furthermore, we discussed the presence and the importance of a finite-size effect: the maximum number of links occurs away from the critical value of the system parameter. We derive an exact Master Equation for this model using a quantum algebra approach to stochastic processes [2]. In the mathematical derivation, fluctuations are much more important than the mean-field approximation predicts, and we attribute it to the heterogeneity in the model. The maximal heterogeneous population corresponds to the critical value of the system where we observe the strongest presence of the emergent properties.

As an economic scenario, we introduce a minimal framework of a financial market where actors are banks which are of two types. Each group has a different (i.e.

opposite) perceptions of risk: high-target leverage banks which are the generators of interbank links, and low-target leverage banks which on the contrary tend to reduce their connections. The borrowing banks will contact randomly different neighboring lending banks. The leveraging actions allows borrowers to purchase more assets then their wealth otherwise permit. The expected returns from the assets represents the possible profit of the leveraging institutions. We show how the critical regime is the most uncertain but profitable state among the different heterogeneity levels. Beyond that, we include a sketched endogenous mechanism to make the heterogeneity parameter of the population to vary over time according to economic relation between the asset price and the number of investors.

2 Results

In order to derive an equation of motion for the link distribution, we write a Fokker-Planck equation for this model starting from the formalism of creation and annihilation operators. The only free parameter in this representation is the total rate of events. In [2] we find the expected number of links according to the heterogeneity variable $\varDelta$ which represent the difference between the two groups as $\varDelta = \frac{N_1 - N_0}{N}$ where $N = N_0 + N_1$ is the total number of units.

The network connectance (link density) is related to the ratio between the average number of actual links over the potential links

$$\ell = \frac{2L}{N_1 N_0} - 1 \,, \qquad \ell \in [-1, 1] \tag{1}$$

The analytical equation for the connectance is the sigmoid function:

$$\langle \ell \rangle_{eq.} = \begin{cases} \dfrac{\sqrt{\varDelta^2 + (2/N)^2} - 2/N}{\varDelta} & \text{for} \quad \varDelta \neq 0 \\ 0 & \text{for} \quad \varDelta = 0, \end{cases} \tag{2}$$

Additionally, we highlight some limitations of the mean field approximation in capturing the heterogeneity of the nodes in their dynamics of creating and destroying links. These corrections are non-negligible when the system is at its critical point as pointed out in Fig.2, where we show that the degree correlations among nodes is not-negligible for values of heterogeneity close to the critical value.

We set forth an abstract example of a system which gains value according to its interconnectedness, and bears a cost depending on the number of active nodes (i.e., generators of links). The resulting profitability shows a signature of complexity in terms of finite-size network effects: small groups reach a maximal profitability far from the critical point of maximal heterogeneous population, but they tend to suffer less uncertainty of the expected connectivity.

In Fig.3 we summarize these results; here the system is profitable inside an interval where the network has varying uncertainty. In the time series perspective, the system can end up in long time periods during which it is exposed to losses, and the whole network could become more susceptible to losses, triggering systemic crisis and possible collapses.

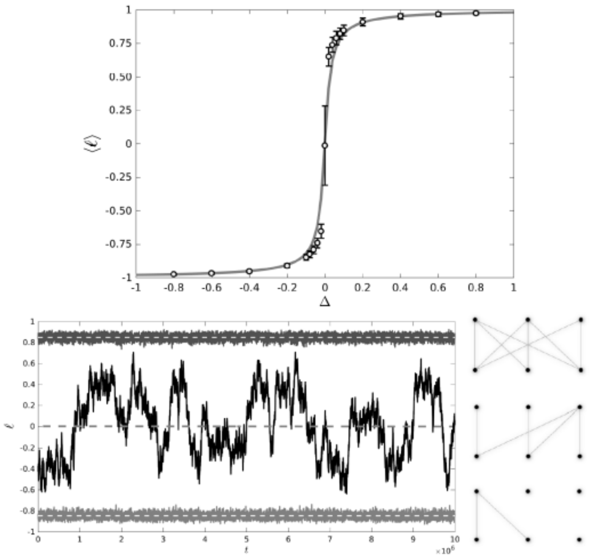


Fig. 1. Emergent properties of the network. Top figure represent the sigmoidal phase transition curve of the expected connectivity vs the heterogeneity population parameter Δ. The bottom figure represents three different cases of Δ the central time series correspond to the critical point $\Delta = 0$ where we observe anomalously large fluctuations and abrupt changes passing from sparse network to dense one in a single trajectory. The top and the bottom figures are the super-critical $\Delta > 0$ and sub-critical case $\Delta < 0$, where we observe poissonian fluctuations.

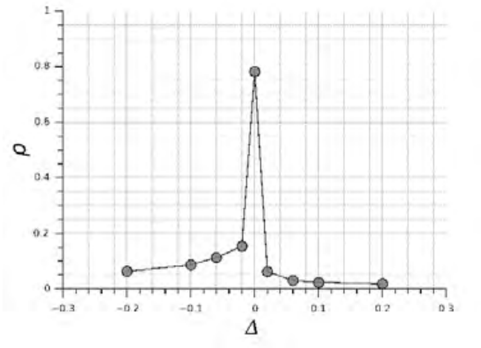


Fig. 2. Average node-degree correlation. It is evident how at criticality the correlation is much higher then in the off-critical cases. This shows how neglecting the correlation among the nodes make the master equation approach fail at criticality.

In the application of financial markets, minimizing the importance of heterogeneity also in mathematical terms (by using the mean-field approximation) leads one to drastically underestimate the size of fluctuations at the critical point, which could lead to an underestimation of the instability of the system. We investigate how shocks on external assets can be absorbed or propagate through the network for different values of hetero-

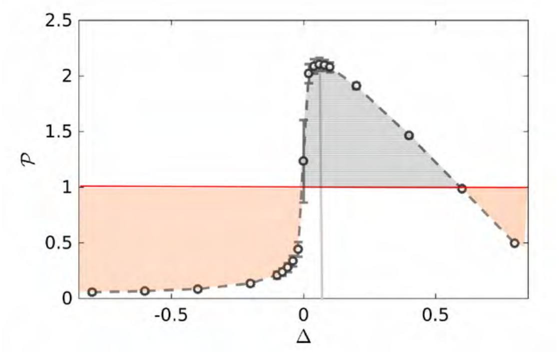


Fig. 3. Profitability is defined as the benefit proportional to the actual number of created links over the cost which is proportional to the number of generators. The gray shaded area is the condition of profitability for the network. The error bar is the uncertainty derived to the the fluctuations of the link formation.

geneity parameter. According to cascades of deleveraging actions, these shocks induce different intensities of the global failures for different levels of the network interconnectivity .

In conclusion, we described a different linking scheme which generates a natural evolving network with just few ingredients show complex behaviors in structural and temporal characteristics. According to certain levels of heterogeneity we characterized a system fragility and resilience where connectivity fluctuations play a key role in terms of temporal and node-degree correlations.

References

1. Liu, Wenjia and Schmittmann, Beate and Zia, RKP, Extraordinary variability and sharp transitions in a maximally frustrated dynamic network, EPL (Europhysics Letters), 100(6), 2013.
2. Lambert D. , Vanni F. : Complexity and Heterogeneity in a Dynamic Network. LEM working paper series, 2017/21 (2017) submited to Chaos, Solitons & Fractals.
3. Vanni F. , Barucca P.: Time evolution of an agent-driven network model. LEM working paper series, 2017/16 (2017) submitted to Complexity.

Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes

Owen T. Courtney[1] and Ginestra Bianconi[1]

School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London
E1 4NS, UK
o.t.o.courtney@qmul.ac.uk

Network theory has been successful over the last fifteen years in characterising social, technological and biological networks. Nevertheless, the increasingly large data sets available in the field require the development of more sophisticated models of networks such as multilayer networks and generalised network structures [1]. In particular a wide variety of networks, including collaboration networks, protein interaction networks and folksonomies, can be modelled by simplicial complexes.

Simplicial complexes are a generalisation of networks constructed using not only nodes and links (that are respectively simplices of dimension zero and one) but also using triangles (simplices of dimension $d = 2$), tetrahedra (simplices of dimension $d = 3$) and higher dimensional simplices. An example of a simplicial complex with dimension $d = 2$ is illustrated in Figure 1a.

Simplicial complexes allow for the treatment of networks with many short loops, and therefore constitute a way to go beyond the current limitation of most statistical mechanics approach to networks that are restricted to locally tree-like networks. At the same time, simplicial complexes, might be studied from the perspective of network geometry, and as such have been proposed to characterise the underlying emergent geometry of complex networks [2].

Here we will present recent results [3] on the configuration model for simplicial complexes of dimension d. This model can be used as a null model to which real networks can be compared and as a benchmark to run simulations of dynamical processes. The configuration model of simplicial complexes of dimension d generalises on one side the configuration model of networks with given degree sequence, and on the other the uncorrelated hypergraph model proposed in [4].

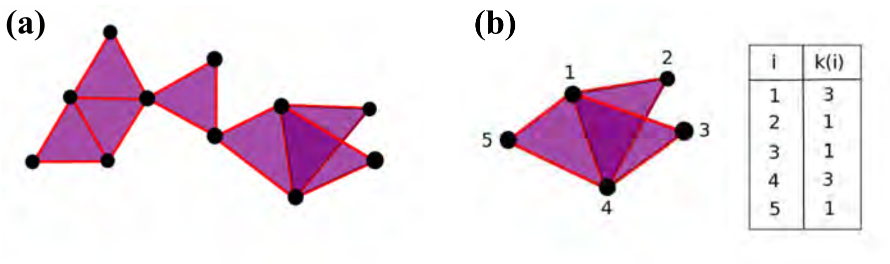
The configuration model of simplicial complexes is formed by all the complexes that have a given sequence of the generalised degree of the nodes, indicating the number of simplices incident to each node (see Figure 1b for an example of generalized degree sequence).

Similarly to the case of the configuration model of networks with a given degree sequence, the configuration model for simplicial complexes can develop natural correlations if there is no structural cutoff on the generalized degrees of the nodes.

Interestingly the structural cutoff for simplicial complexes depends on the dimension d of the simplicial complex, and for dimension $d > 1$ the cutoff value differs from the that found for simple networks.

Our work includes a fully statistical mechanical characterization of the configuration model of simplicial complexes. In particular we are able to compare properties of this ensemble with those of its conjugated canonical ensemble, which corresponds to an Exponential Random Simplicial Complex with given expected generalized degrees of the nodes.

We are able to evaluate and compare the entropy of these two ensembles. The entropy evaluates the logarithm of the number of typical simplicial complexes belonging to each of the ensembles. This study shows that the two ensembles are not asymptotically equivalent, and opens up new scenarios for solving inference problems on simplicial complexes.



i	k(i)
1	3
2	1
3	1
4	3
5	1

Fig. 1. Panel (a): Example of a simplicial complex with dimension $d = 2$ on 11 nodes. Panel (b): Example illustrating the generalised degree of the nodes for a simplicial complex with dimension $d = 2$. The table gives the generalised degree $k(i)$ for each vertex with label i, given by the number of triangles incident to node i .

References

1. Bianconi, G.: Interdisciplinary and physics challenges of Network Theory. Europhys. Lett. 111 56001 (2015)
2. Wu, Z., Menichetti, G., Rahmede, C., Bianconi, G.: Emergent complex network geometry. Scientific Reports 5, 10073 (2015)
3. Courtney, O.T., Bianconi G.: Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. Phys. Rev. E 93, 062311 (2016)
4. Ghoshal, G., Zlati, V., Caldarelli, G., Newman, M.E.J.: Bond percolation on a class of correlated and clustered random graphs. Phys. Rev. E 79, 066118 (2009)

Machine learning meets complex networks via coalescent embedding of networks in the hyperbolic space

Alessandro Muscoloni[1], Josephine Maria Thomas[1], Sara Ciucci[1,3], Ginestra Bianconi[4] and Carlo Vittorio Cannistraci[1,2,*]

[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[2] Brain bio-inspired computation (BBC) lab, IRCCS Centro Neurolesi "Bonino Pulejo", Messina, Italy
[3] Lipotype GmbH, Tatzberg 47, 01307 Dresden, Germany
[4] School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom

The Popularity Similarity Optimization (PSO) model suggests that the trade-off between node popularity and similarity explains how complex network topologies emerge, as discrete samples, from the continuous world of hyperbolic geometry [1]. On one hand, node similarities are related with the angular distances, on the other hand, the node degree is related with the intrinsic popularity of the node. The hyperbolic space, indeed, preserves many of the fundamental topological properties of real complex networks. Hence, one of the most challenging problems of recent complex network theory is to map a given network to its hyperbolic space.

Manifold machine learning for unsupervised nonlinear dimensionality reduction is an important sub-class of topological machine learning algorithms. They learn nonlinear similarities between points distributed over a hidden manifold in a multidimensional feature space, in order to preserve, embed and visualize them in a reduced space [2], [3].

Here, adopting different techniques, we show that the node angular coordinates of the hyperbolic model can be directly approximated according to a similar geometrical node aggregation pattern that we name *angular coalescence* (Fig. 1). Based on this phenomenon, we propose a class of algorithms that offers fast and accurate *coalescent embedding* in the hyperbolic space even for large networks.

The methods outperform the state-of-the-art for accuracy of mapping in the hyperbolic space and, at the same time, reduce the computational complexity from O(N^3)-O(N^4) of current techniques [4], [5] to O(N^2) (Fig. 2). Furthermore, the algorithms can embed also weighted networks and in hyperbolic spaces of two or more dimensions. Several studies can be lead exploiting the geometrical information and this achievement can have an impact for many disciplines including biology, medicine, computer science and physics.

References

[1] F. Papadopoulos, M. Kitsak, M. A. Serrano, M. Boguna, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, pp. 537–540, 2012.

[2] C. V. Cannistraci, T. Ravasi, F. M. Montevecchi, T. Ideker, and M. Alessio, "Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes," *Bioinformatics*, vol. 26, pp. i531–i539, 2010.

[3] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, "Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding," *Bioinformatics*, vol. 29, no. 13, pp. 199–209, 2013.

[4] F. Papadopoulos, C. Psomas, and D. Krioukov, "Network mapping by replaying hyperbolic growth," *IEEE/ACM Trans. Netw.*, vol. 23, no. 1, pp. 198–211, 2015.

[5] F. Papadopoulos, R. Aldecoa, and D. Krioukov, "Network Geometry Inference using Common Neighbors," *Phys. Rev. E*, vol. 92, no. 2, p. 22807, 2015.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction.," *Science*, vol. 290, pp. 2319–23, 2000.

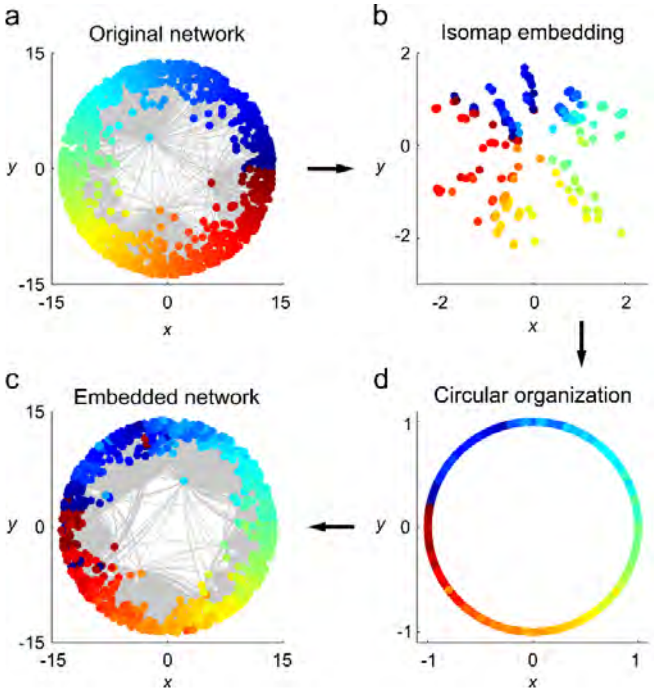


Fig. 1. (a) We show the original synthetic network generated by the PSO model in the hyperbolic space. (b) The Isomap algorithm (ISO) [6], which is the progenitor of manifold-based techniques, starting from the unweighted adjacency matrix offers an embedding of the network nodes that is organized according to a circular pattern that follows the angular coordinates of the original PSO model. (d) The angular coordinates are given projecting the nodes over the circumference and adjusting them equidistantly. (c) Assigning also the radial coordinates, according to a mathematical formula which is function of the power-lawness of the degree distribution, the final embedded network is obtained.

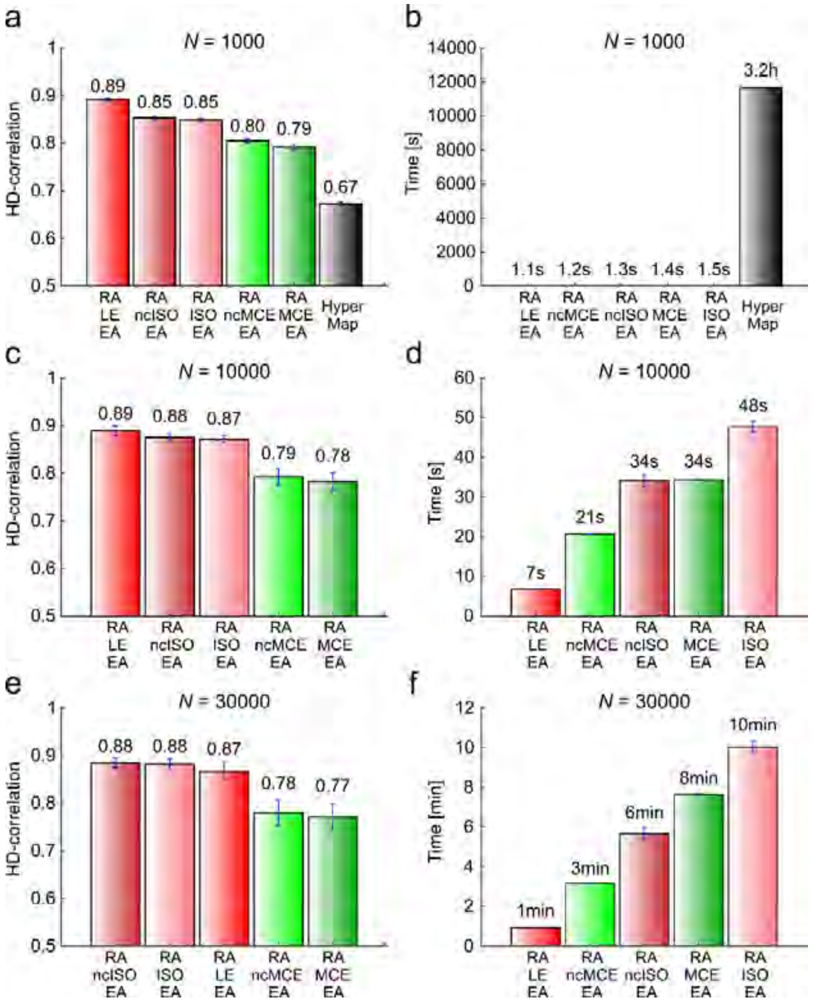


Fig. 2. (a, c, e) Correlation between original and inferred pairwise node hyperbolic distances (HD-correlation), which measures the mapping accuracy. The plots report for increasing network size N = [1000, 10000, 30000] the average and standard error over several combinations of the PSO model parameters m (half of the mean node degree) and T (temperature, inversely related to clustering). (b, d, f) For the same parameter combinations the average computational time is reported. Considering the average HD-correlation on 1000 nodes networks (a), coalescent embedding approaches achieved a performance improvement of more than 30% in comparison to the state-of-the-art method HyperMap, requiring only around one second versus more than three hours of computation time. Similar performance results are confirmed for the networks of sizes N = [10000, 30000] with an execution time still in the order of minutes for the biggest networks. The comparison to HyperMap was not possible due to its long running time.

A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities

Alessandro Muscoloni[1] and Carlo Vittorio Cannistraci[1,2,*]

[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[2] Brain bio-inspired computation (BBC) lab, IRCCS Centro Neurolesi "Bonino Pulejo", Messina, Italy

The hidden metric space behind complex network topologies is a fervid topic in current network science and the hyperbolic space is one of the most studied, because it seems associated to the structural organization of many real complex systems. The Popularity-Similarity-Optimization (PSO) model [1] simulates how random geometric graphs grow in the hyperbolic space, reproducing strong clustering and scale-free degree distribution, however it misses to reproduce an important feature of real complex networks, which is the community organization. The Geometrical-Preferential-Attachment (GPA) model [2] was recently developed to confer to the PSO also a community structure, which is obtained by forcing different angular regions of the hyperbolic disk to have variable level of attractiveness. However, the number and size of the communities cannot be explicitly controlled in the GPA, which is a clear limitation for real applications. Here, we introduce the nonuniform PSO (nPSO) model that – differently from GPA - forces heterogeneous angular node attractiveness by sampling the angular coordinates from a tailored nonuniform probability distribution, for instance a mixture of Gaussians (Fig. 1). The nPSO differs from GPA in other three aspects: it allows to explicitly fix the number and size of communities; it allows to tune their mixing property through the network temperature; it is efficient to generate networks with high clustering. After several tests we propose the nPSO as a valid and efficient model to generate networks with communities in the hyperbolic space, which can be adopted as a realistic benchmark for different tasks such as community detection (Fig. 2) and link prediction.

References

[1] F. Papadopoulos, M. Kitsak, M. A. Serrano, M. Boguna, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, pp. 537–540, 2012.

[2] K. Zuev, M. Boguna, G. Bianconi, and D. Krioukov, "Emergence of Soft Communities from Geometric Preferential Attachment," *Sci. Rep.*, vol. 5, p. 9421, 2015.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech Theory Exp.*, vol. 2008, no. 10, p. 10008, 2008.

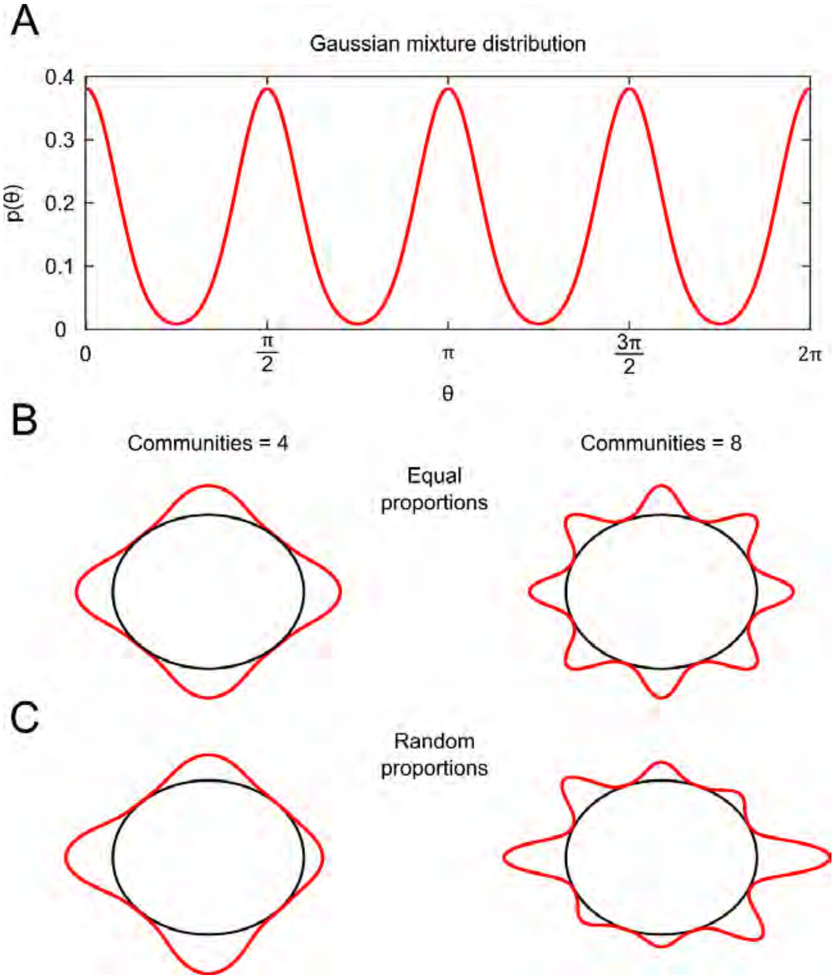


Fig. 1. The figure shows examples of nonuniform distributions used for sampling the angular coordinates of the nodes. The distributions are generated using a Gaussian mixture model, with as many components as the number of the desired communities, placing the mean of the components equidistantly over the angular space and with equal standard deviations, whose value is chosen as 1/6 of the distance between two adjacent means. (A) Plot of the Gaussian mixture distribution using 4 components having the same mixing proportion. (B) Representation of the Gaussian mixture distribution, using 4 and 8 components having the same mixing proportion, along the angular space of the hyperbolic disk. (C) Representation of the Gaussian mixture distribution, using 4 and 8 components having random mixing proportions, along the angular space of the hyperbolic disk.

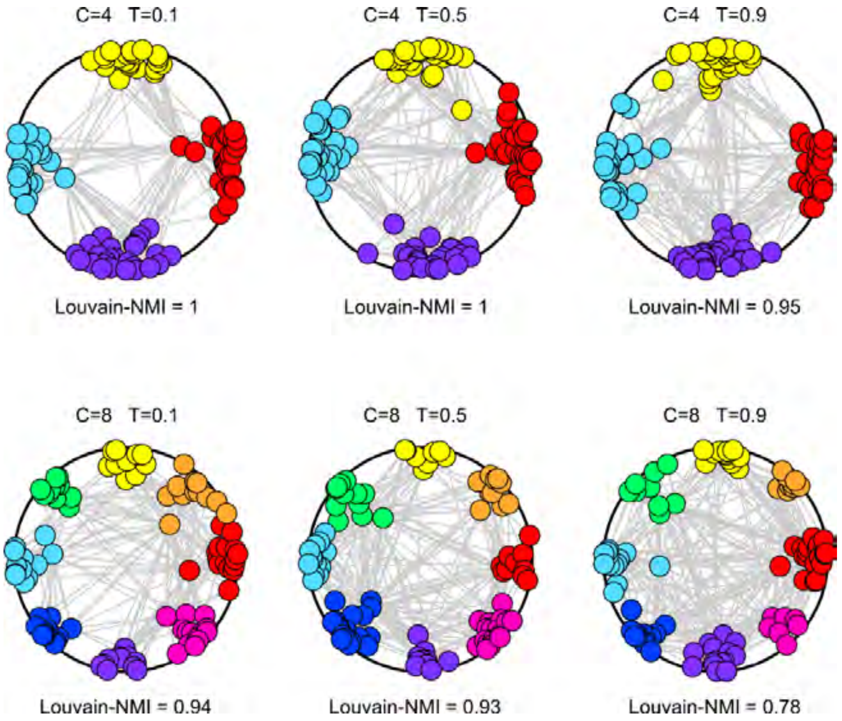


Fig. 2. Synthetic networks have been generated using the nPSO model with parameters $\gamma = 3$ (power-law degree distribution exponent), $m = 5$ (half of average degree), $T = [0.1, 0.5, 0.9]$ (temperature, inversely related to the clustering coefficient), $N = 100$ (network size) and $C = [4, 8]$ (communities). For each combination of parameters, 10 networks have been generated. For each network the Louvain community detection method [3] has been executed and the communities detected have been compared to the annotated ones computing the Normalized Mutual Information (NMI). The plots show for each parameter combination a representation in the hyperbolic space of the network that obtained the highest NMI, whose value is reported. The nodes are coloured according to the communities as generated by the nPSO model. We notice that the communities are perfectly detected both for C = 4 and C = 8 at low temperature, suggesting that a meaningful community structure is generated by the proposed model. For the same number of communities, if the temperature is increased the performance slightly decreases, because more inter-community links are established in the network, causing as expected higher rate of wrong assignments by the community detection algorithm.

Raising Graphs from Randomness

Róbert Pálovics   András A. Benczúr   Ferenc Béres

Institute for Computer Science and Control of the
Hungarian Academy of Sciences (MTA SZTAKI),
{palovics,benczur,beres.ferenc}@sztaki.hu

1 Introduction

In our work we analyze the fine-grained connections between the average degree and the power-law degree distribution exponent in growing information networks. Our starting observation is a power-law degree distribution with a decreasing exponent and increasing average degree as a function of the network size. Our experiments are based on three Twitter at-mention networks and three more from the Koblenz Network Collection. We observe that popular network models cannot explain decreasing power-law degree distribution exponent and increasing average degree at the same time.

We propose a model that is the combination of exponential growth, and a power-law developing network, in which new "homophily" edges are continuously added to nodes proportional to their current homophily degree. Specifically, we connect the growth of the average degree to the decreasing exponent of the power-law degree distribution. Prior to our work, only one of the two cases were handled. Existing models and even their combinations can only reproduce some of our key new observations in growing information networks.

Our finding appeared as

- Róbert Pálovics and András Benczúr. Raising graphs from randomness to reveal information networks. In *Proceedings of The 10th ACM International Conference on Web Search and Data Mining*, 2017.

2 Our results

We study the growth of information networks by considering processes where each node and edge is added to the network only once, and no node or edge is deleted from the network. Our key finding is related to how the average degree and the power-law degree distribution of the network evolve over time. More specifically,

- the exponent of the power-law degree distribution in the network *decreases down to two* over time, and
- the average degree grows as $a + cn^b$, where n is the number of nodes in the network.

For example, as seen in Figure 1 left, in graphs generated by the Barabási-Albert model [1], the degree distribution exponent stays very close to constant. In contrast in our measurements the degree distribution log-log plot lines of real networks get flattened (see Fig. 1 right).

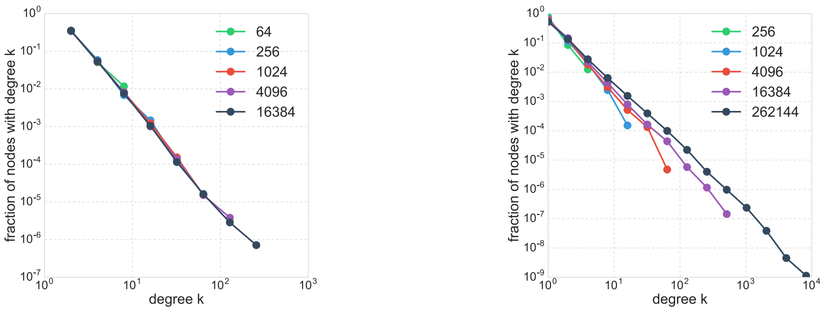


Fig. 1. Degree distribution snapshots of growing networks at different sized (number of nodes) indicated in the legend. **Left:** The Barabási-Albert model yields fixed exponent. **Right:** The Occupy Twitter mention data set with flattening slope as the network grows.

We emphasize the importance of the constant a in the average degree formula. The constant was considered negligible in the experiments of Leskovec et al. [2]. In our results, however, the constant helps to capture the mixture of edges that appear at random vs. as a result of common interest, and fit to the actual measurements (see Fig. 2, left).

To our best knowledge, there is no graph model yet that captures all the observed effects simultaneously. Leskovec et al. [2] observe densification, a power-law growth for the average degree. Their models apply to graphs where the exponent of the degree distribution is less than 2 and remains constant over time. They predict densification for networks with power law exponent larger than 2 that is the case for all of our real networks, however they give no models.

We introduce our new model that relies on preferential attachment, and can generate growing networks with decreasing power law exponents. In our growing network model we add at each time step: (i) random new edges that connect two new nodes in the network, (ii) and edges between already existing nodes in the network. More specifically, at time t, when the number of nodes is $n(t)$ in the network (see Figure 2, right):

- For some constant r, $r \cdot n(t)$ new *random edges* appear that indicate the random growth of the network.
- Each node i selects other nodes to connect with *homophily edges* randomly. The expected number of new homophily connections created by node i is $s \cdot d_h(i)$, where $d_h(i)$ is the number of homophily edges already connected to node i, i.e. the homophily degree of node i. For a given new connection of node i, the target node is selected by preferential attachment. In other words, the probability of selection for node j as a new neighbor of i is the degree of j $d(j)$.

The main difference of our model compared to earlier models can be summarized in three points.

- The power law exponent, as in all our real networks, is greater than 2, this could not be modeled in [2].

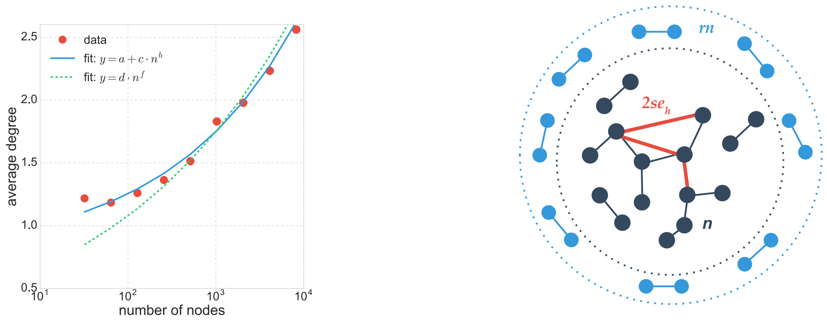


Fig. 2. Left: Growth of the average degree in the Occupy Twitter mention data set. **Right:** Visualization of our proposed model: at each time t, we add $rn(t)$ random, and $2se_h(t)$ preferential edges.

- Our model explains the initial behavior of the degrees as a natural mixture of influence and preferential attachment edges, and also predicts correctly the ratio of these edges.
- Our model generates both increasing average degree and decreasing power law exponent.

3 Summary

In evolving networks, we measured that the exponent of the power-law degree distribution decreases in time. We connected our observation on decreasing exponent with the growth of the average degree and gave models for these two phenomena.

As a general overview of the possible models based on our observations, networks start to grow at random, like an Erdős-Rényi graph. Then certain rules such as preferential attachment [1] intensifies during the growth process, and causes scale-free degree distribution with a decreasing exponent. The stronger the rule is, the closer the exponent of the degree distribution gets down to two in a more coherent network. As the degree distribution log-log plot flattens, the chance for very high degree nodes in a strongly skewed distribution increases acting as the main organizer of the network structure.

As our main result, we model the transition from a random and mostly disconnected graph to a highly organized and very skewed degree distribution network. Our model is based on exponential growth and preferential attachment. The model yields a power-law degree distribution with decreasing exponent in a growing graph with increasing average degree.

References

1. Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
2. Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

Analytical expression for the size distribution of connected components in the infinite configuration model

Ivan Kryven[1]

University of Amsterdam, Amsterdam 1090 GE, The Netherlands,
i.kryven@uva.nl,
home page: staff.fnwi.uva.nl/i.kryven

1 Introduction

In the configuration network, N nodes are assigned predefined degrees. The edges connecting these nodes are then considered to be random, and every distinct configuration of edges that satisfies the given degree sequence is treated as a new instance of the network in the sense of random graphs. When the number of nodes N approaches infinity, the degree sequence, which provides the only input information for the model, is equivalent to the frequency distribution of degrees, $u(k), k = 1, 2, ...$, i.e., the probability that a randomly chosen node has degree k. Size distribution of connected components, $w(n)$, denotes probability that a randomly chosen node is part of a connected component of finite size n. Connected components in the infinite configuration network can be of finite or infinite size. Depending upon a specific context behind the network, the size distribution of connected components may summarise important features of the modelled system, for instance as it often happens in chemistry [4, 3], where connected components represent irregular molecular structures; epidemiology [5], where configuration model is widely used in modelling disease outbreaks; or linguistics [6], where connected components are proved to be useful when studying sentence similarity graphs and structure of natural languages. This brief list of application case studies is far from being exhaustive.

Let $U(x)$ denotes the generating function (GF) of the degree distribution, $U(x) = \sum_{k=0}^{\infty} u(k)x^k$, $x \in \mathbb{C}$, $|x| \leq 1$. Newman et al. [7] showed that the GF of the degree distribution can be put into a correspondence to the GF of the size distribution of connected components $W(x)$. Namely, it has been shown that $W(x)$ satisfies the following system of functional equations

$$W(x) = xU[W_1(x)], \tag{1}$$

$$W_1(x) = xU_1[W_1(x)], \tag{1'}$$

where $U_1(x)$ is the GF for the excess degree distribution $u_1(k) = \frac{k+1}{\mu_1}u(k+1)$ and μ_1 is the expected degree of a node. Many studies refer to the *numerical solution* of eq. (1) as the method of choice for recovering the size distribution of connected components.

2 Results

By exploiting tools from analytical combinatorics, we have showed in Ref [1] that Eq. (1) can be solved analytically for arbitrary degree distributions featuring a finite first moment, $\mu_1 = \sum_{k=1}^{\infty} ku(k)$. Namely, the probability that a randomly sampled node belongs to a connected component of size n is given by

$$w(n) = \begin{cases} \frac{\mu_1}{n-1} u_1^{*n}(n-2), & n > 1, \\ u(0) & n = 1. \end{cases} \tag{2}$$

where $u_1^{*n}(k) = u_1(k) * u_1^{*n-1}(k)$ is the n-fold convolution of $u_1(k)$. In fact, probability mass function $w(n)$ is an analytical function: the first five values of $w(n)$ read as,

$$\begin{aligned} w(1) =& u(0), \\ w(2) =& \frac{1}{\mu_1} u(1)^2, \\ w(3) =& \frac{3}{\mu_1^2} u(1)^2 u(2), \\ w(4) =& \frac{4}{\mu_1^3} u(1)^2 [2u(2)^2 + u(1)u(3)], \\ w(5) =& \frac{5}{\mu_1^4} u(1)^2 [4u(2)^3 + 6u(1)u(2)u(3) + u(1)^2 u(4)], \\ & \dots \end{aligned}$$

The number of terms in the expansion (2) is rapidly growing with n, nevertheless, it is possible to compute exact value of $w(n)$ by spending not more then $O(n \log n)$ multiplicative operations by applying smart reorganisation of the terms. A graphical comparison of the theoretical predictions and simulated data can be seen in Fig. 1. Furthermore, Eq. (2) is tractable from the point of view of asymptotic theory. The asymptotical analysis is based on the fact that $n-$fold convolution can be interpreted as a probability for a one-dimensional random walk to hit a specific value in n steps and results in a plethora of asymptotic modes, many of which have not been documented before, see Table 1. These asymptotical modes include exponential and scale free asymptotes, but also 'exotic' asymptotic cases of faster-then-algebraic but slower-then-exponential decays. Additionally, we highlight existence of 'transient' asymptotical modes, that are not asymptotes in the strict mathematical sense but may appear to look like asymptotes when when limited amount of data is analysed.

Summary. This work presents a simple equation that gives for an arbitrary degree distribution the corresponding size distribution of connected components. This equation is suitable for fast and stable numerical computations up to the machine precision. The analytical analysis reveals that the asymptote of the component size distribution is completely defined by only a few parameters of the degree distribution: the first three moments, scale, and exponent (if applicable). When the degree distribution features a heavy tail, multiple asymptotic modes are observed in the component size distribution that, in turn, may or may not feature a heavy tail.

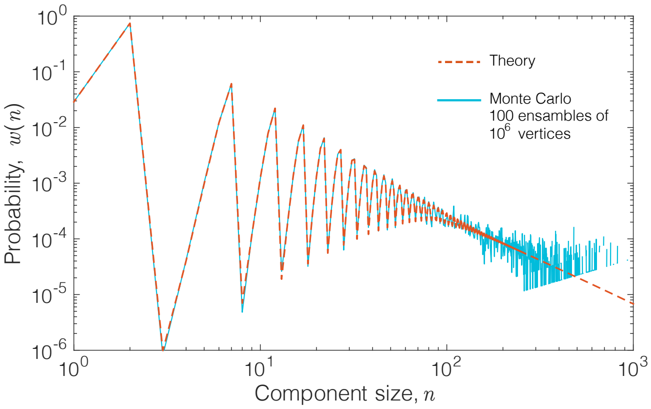


Fig. 1. Comparison of the theoretically predicted size distribution of connected components against Monte Carlo generated data.

Finite moments of $u(k)$	$u(k)$, $k \to \infty$	$\theta = \mu_2 - 2\mu_1$	Asymptote of $w(n)$
$\mu_3 < \infty$	A. $o(k^{-\beta}),\ \beta > 4$	$\theta \neq 0$	$C_1 e^{-C_2 n} n^{-3/2}$
		$\theta = 0$	$C_1 n^{-3/2}$
	B. $O(k^{-\beta}),\ \beta > 4$	$\theta < 0$	$C_3 n^{-\alpha-1}$
		$\theta = 0$	$C_1 n^{-3/2}$
		$\theta > 0$	$C_1 e^{-C_2 n} n^{-3/2}$
$\mu_3 = \infty,\ \mu_2 < \infty$	C. $O(k^{-\beta}),\ \beta = 4$	$\theta < 0$	$C_3 n^{-\alpha-1}$
		$\theta = 0$	$C_1' \frac{n^{-3/2}}{\sqrt{\log n}}$
		$\theta > 0$	$C_1' \frac{n^{-3/2}}{\sqrt{\log n}} e^{-C_2' \frac{n}{\log n}}$
	D. $O(k^{-\beta}),\ 3 < \beta < 4$	$\theta < 0$	$C_3 n^{-\alpha-1}$
		$\theta = 0$	$C_4 n^{-\frac{1}{\alpha}-1}$
		$\theta > 0$	$C_5 e^{-C_6 n} n^{-3/2}$
$\mu_2 = \infty$	E. $O(k^{-\beta}),\ \beta = 3$	$\theta > 0$	$C_7 e^{-C_8 - C_9 n^{\frac{2}{\pi}}} n^{\frac{1}{\pi}-2}$
	F. $O(k^{-\beta}),\ 2 < \beta < 3$	$\theta > 0$	$C_{10} e^{-C_{11} n} n^{-3/2}$

Table 1. Asymptotic behaviour of the size distribution of connected components in terms of degree distribution parameters: the first three moments μ_1, μ_2, μ_3, scale parameter s, and exponent β. In the expressions of the asymptotes, $\alpha = \beta - 2$ and the values of constants $C_{1,\dots,9}$ are as given in Ref [1].

References

1. I. Kryven: General expression for the component size distribution in infinite configuration networks. Physical Review E 95.5 (2017)
2. M. Molloy and B. Reed, Random structures & algorithms 6, 161 (1995)
3. I. Kryven, Journal of Mathematical Chemistry, 10.1007/s10910-017-0785-1 (2017).
4. V. Schamboeck, I. Kryven, and P. Iedema, MTS, 10.1002/mats.201700047 (2017)
5. M. E. Newman, D. J. Watts, and S. H. Strogatz: PNAS 99, 2566 (2002).
6. C. Biemann and A. van den Bosch: Structure discovery in natural language (2011).
7. M. E. J. Newman, S. H. Strogatz, and D. J. Watts: Physical review E 64, 026118 (2001).

Multiplex Network Regression: How Do Relations Drive Interactions?

Giona Casiraghi

ETH Zürich, Chair of System Design, Weinbergstrasse 56/58, 8092 Zürich, Switzerland,
giona@ethz.ch

1 Introduction

We often deal with datasets of *observed repeated interactions* between elements of a system. These datasets are used to generate networks where the elements are represented by vertices and interactions by edges. We ask whether these interactions are random events or whether they are driven by existing relations between the elements. To answer this question, we propose a statistical model to regress *relations*, which we identify as *independent variables*, on a network created from interactions, which we will refer to as *dependent variables*.

In general, a regression model explains dependent variables as a function of the independent ones, accounting for random effects. Parallel to linear regression models, which model the relationship between a dependent variable and multiple explanatory variables, here we assume that the observed interactions are driven by different relations, masked by *combinatorial effects*. With combinatorial effects, we mean that elements that interact more in general are also more likely to interact with each other, even if they have no relations. This problem is well known in network theory, where it is referred to as *degree-correction* (see e.g., [5, 4, 2]). For example, the fact that two individuals have contact very often can be explained by multiple reasons. They may interact because they are friends, because they work together, or simply because they are very active, and hence have high chances to meet. Therefore, to have a full understanding of the system, we have to disentangle relations from combinatorial effects.

There exist few statistical models addressing the problem of *quantifying* the interdependence between observed edges and dyadic relations. This problem is exacerbated by the fact that the dyadic relations represented in complex networks are not independent from one another. We can summarize the limitations of existing methods into two main issues. First, many of them are not appropriate for weighted graphs. Second, they do not take into account the combinatorial effects typical of interaction data.

2 Results

To solve these issues, we propose a new nonlinear parametric model to perform statistical regression on networks. Our method is based on an extension of *generalized hypergeometric ensembles* – a class of analytically tractable ensembles for weighted directed networks we recently developed [1] – to multiplex applications. Generalized hypergeometric ensembles contain random graphs generated by merging arbitrary relations between vertices and combinatorial effects. This model allows to regress the influence of each relational layer, i.e. the independent variables, on the interaction counts, i.e. the dependent variable .Additionally, we will show how to quantify the significance of the computed parameters and the goodness-of-fit of the model.

We want to model the interaction layer $\mathscr{I}$, which is a multi-edged graph with fixed number of edges m. To do so, we treat $\mathscr{I}$ as a realization from a *generalized hypergeometric ensemble* $\mathbb{E}(n,m)$, with n vertices and m edges. We indicate with $\mathbf{A}$ the adjacency matrix of the interaction layer $\mathscr{I}$ and A_{ij} with $i,j \in V$, its elements. Similarly, let $\mathbf{R}_l$ be the adjacency matrix capturing the known relations between vertices, corresponding to the relational layer $\mathscr{R}_l$, and with $\beta \in \mathbb{R}^r$ the r-vector of regression coefficients. $\mathscr{I}$ is then distributed according to the Wallenius non-central hypergeometric distribution [6, 1]

$$\Pr(\mathscr{I}|\mathscr{R}) = \left[\prod_{i,j}\binom{\Xi_{ij}}{A_{ij}}\right]\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}}\right)^{A_{ij}} dz \qquad (1)$$

with $S_\Omega = \sum_{i,j}\Omega_{ij}(\Xi_{ij} - A_{ij})$.

The distribution in eq. (1) is defined by the two quantities Ξ and Ω. Ω encodes the tendency of pairs of vertices to connect beyond combinatorial effects, and Ξ the probability that pairs of vertices are connected because of combinatorial effects. For simplicity we assume the entries of the matrix of possible edges Ξ are built according to the configuration model. This is the most general way to encode combinatorial effects generated by the different activity, i.e., degree, of vertices. It means that vertices that are more active, i.e., have higher degree, are more likely to interact. Hence, Ξ is completely defined by $\mathscr{I}$. On the other hand, Ω depends on the relational layers $\{\mathscr{R}_l\}_{l\in[1,r]}$ as follows:

$$\Omega := \prod_{l=1}^{r} \mathbf{R}_l^{\beta_l}. \qquad (2)$$

We can hence specify a statistical model as follows:

$$\mathscr{I} = \mu\,[\hat{\mathbb{E}}(n,m)|\mathscr{R}_1,\ldots,\mathscr{R}_r], \qquad (3)$$

where $\mathscr{I}$ is set to be equal to the expectation of the generalized hypergeometric ensemble defined by the relational layers $\mathscr{R}_i$. We estimate the model in eq. (3) by finding the maximum likelihood estimators (MLE) for the parameter vector β in eq. (1).

To demonstrate the power of our approach and its broad applicability, we present examples based on synthetic and empirical data. Many datasets including both interactional and relational information exist. Networks built from face-to-face encounters between individuals, supported by underlying social networks and other similar relations (e.g. *SocioPatterns*) fall in this category. As case study, we will use (i) the *US Cosponsorship Network*, where the interaction layer is built from the cosponsorship counts between members of the Congress, along with further relational layers built from party membership, ideological distance, and state provenance; and (ii) the SocioPattern dataset provided in [3]. Here the data available consist of an interaction network, built from recorded contact counts between high-school students, and of further information such as student's gender, class membership and topic, self-reported friendship relations, and Facebook connections.

In both cases, our novel network regression framework is able to identify the relevant factors biasing the interaction layer, and allows to quantify the size of their effects. This method can be applied to a plethora of cases where we want to model counts of interactions between entities.

References

1. Giona Casiraghi, Vahan Nanumyan, Ingo Scholtes, and Frank Schweitzer. Generalized Hypergeometric Ensembles: Statistical Hypothesis Testing in Complex Networks. *arXiv preprint arXiv:1607.02441*, jul 2016.
2. Brian Karrer and M E J Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83(1):16107, 2011.
3. Rossana Mastrandrea, Julie Fournet, and Alain Barrat. Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE*, 10(9):1–26, 2015.
4. Mark E. J. Newman. Generalized Communities in Networks. *Physical Review Letters*, 115(8):088701, 2015.
5. Tiago P. Peixoto. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X*, 4(1):011047, 2014.
6. Kenneth T Wallenius. *Biased Sampling: the Noncentral Hypergeometric Probability Distribution*. Ph.d. thesis, Stanford University, 1963.

Coherence of multi-agent networks with reaction time delays

Yongzheng Sun[1] and Wang Li[1]

School of Mathematics, China University of Mining and Technology, Xuzhou 221008, PR China,
yzsung@gmail.com,
WWW home page: http://www.scholat.com/yzsung

1 Introduction

Collective animal behaviour is often modelled by multi-agent networks which assume that each agent alters its behaviour according to signals in its neighbourhood. In recent years, many efforts have been made to understand the mechanisms underlying and ensuring coherence behavior [1]-[5]. A topic of intense ongoing research in the coherence of multi-agent networks is the phenomenon that coherent groups often make suddenly changes in direction. Experiments on desert locusts[1] and glass prawns[6] show that the mean switching time is proportional to the groups density. And the directional switching can not be observed if the density is too large. In Ref.[7], the authors find that the inherent noise can facilitate the coherence in collective swarm motion. In our recent study, we show that the information transmission time delay can also facilitate coherence in collective swarming motion [8]. And it is difficult to see the order directional switching if the time delay are too large. However, the factor that can shorten the mean switching time of moving groups remains unknown.

Here, we investigate the directional switching of multi-agent networks with reaction time delays. Both our analytical and numerical results show that, in spite of the time delays, the group can transit to ordered motion as the the density of group increases. But the mean switching time decreases as the reaction delay increases. Our result implies that, different from the information transmission delays [8], the reaction delays play a destructive role in the ordered directional switching of self-propelled particles group.

2 Results

We consider a group of N individuals moving along a one-dimensional circle, which we identify with the interval $\Omega = [0,1)$ with periodic boundary conditions. Each individual is described by its position, $X_i \equiv X_i(t) \in \Omega$, and velocity, $V_i \equiv V_i(t), i = 1,2,\ldots,N$. The ring-shaped domain has a sufficient width that individuals can pass each other, i.e. one-dimensional modelling implicitly assumes that individuals can cross through each other [1,7]. Each individual adjusts its behavior according to the behavior of its neighbours, which can be found less than a distance R (the interaction radius) from it. The time evolution of X_i and V_i is then given by the following equations:

$$\mathrm{d}X_i = V_i(t)\,\mathrm{d}t, \tag{1}$$

$$\mathrm{d}V_i = \Big[\mathrm{sign}\,(U_{i,R}(t-\tau)) - V_i(t-\tau)\Big]\,\mathrm{d}t + \eta\,\mathrm{d}W_i, \tag{2}$$

where $\tau > 0$ is the reaction time delay corresponding to processing, cognitive, or execution delays, $\eta > 0$ is a parameter, $J_{i,R}(t)$ is the set of neighbours of the i-th agent, $\mathrm{d}W_i$ are standard white noise terms (independently sampled for each individual), $\mathrm{sign} : \mathbb{R} \to \{-1,0,1\}$ is the signum function and

$$U_{i,R}(t) = \frac{1}{|J_{i,R}(t)|} \sum_{j \in J_{i,R}(t)} V_j(t) \tag{3}$$

is the mean of the velocities at time t of the individuals within the R-neighbourhood of the i-th individual. We also denote the average velocity of the whole group by $U(t)$, i.e. $U(t) \equiv U_{i,R}(t)$ for R larger than 0.5 and arbitrary i.

Fig. 1. (a) Average velocity $U(t)$ calculated by the model (1)–(3) with $\tau = 0.1, \eta = 2, R = 0.15$, $N = 10$(top panel), $N = 20$(middle panel), and $N = 30$ (bottom panel). (b)Average velocity $U(t)$ calculated by the model (1)–(3) with $N = 20, \eta = 2, R = 0.15$, $\tau = 0.1$(top panel), $\tau = 0.5$(middle panel), and $\tau = 0.9$ (bottom panel).

Let $P(u,t)\mathrm{d}u$ denote the probability that the global velocity average $U(t) \in [u, u + \mathrm{d}u)$. Based on the theory of Fokker-Planck equation, we show that the mean switching time is proportional to the group size N and is inversely proportional to the reaction delays. The analytical results are confirmed by the numerical simulations. We numerically investigate the influence of group density on the collective motion by simulating the model (1)-(3) for different values of N. As the density of individuals increases, we can clearly distinguish two quasi-stationary states (see Fig.**??**) and observe that the group switches suddenly it's velocity to opposite direction(i.e., from left ($U(t) < 0$) to right ($U(t) > 0$), and vice versa). This implies that as the group size N increases, disordered movement of individuals within the group can transit to high aligned collective motion which is agreement with the experimental observations described in [1]. Figure **??** displays the global velocity average $U(t)$ for $N = 30$. We find from Fig.**??** that the mean switching time of group with small reaction delay is longer than that of group with large reaction delay. It is clear that the group can transit to high aligned state as the reaction delay times decreases, which clearly indicates that the reaction time delay can destroy the coherence in collective swarm motion.

Summary. To summarize, we have studied the directional switching of multi-agent networks with reaction time delays. Our results indicated that the reaction time delays can destroy the ordered directional switching behavior. Previous studies have shown that transmission delays can facilitate collective behavior. Future work will extend this work to consider the directional switching behavior of multi-agent networks with both reaction and transmission delays.

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61403393, and 11711530203) and the Fundamental Research Funds for the Central Universities(Grant No. 2017XKZD11)

References

1. Buhl, J., Sumpter, D.J.T., Couzin, I. D., and Simpson, S.J.: From disorder to order in marching locusts. Science 312 (5778), 1402-1406 (2006)
2. Erban, R., Haskovec, J., and Sun, Y.: A Cucker–Smale Model with Noise and Delay. SIAM Journal on Applied Mathematics, 76(4), 1535-1557 (2016)
3. Vicsek, T,, Zafeiris, A.: Collective motion. Physics Reports 517(3), 71-140 (2012)
4. Sun, Y., Li, W., and Zhao, D.: Realization of consensus of multi-agent systems with stochastically mixed interactions. Chaos 26(7), 073112 (2016)
5. Sun, Y., Zhao, D., and Ruan, J.: Consensus in noisy environments with switching topology and time-varying delays. Physica A 389(19), 4149-4161 (2010)
6. Mann, R.P., Perna, A., Strömbom, D., and Ward, A.J.W.: Multi-scale inference of interaction rules in animal groups using Bayesian model selection. PLoS computational biology 9(3), e1002961 (2013)
7. Yates, C. A., Erban, R., Escudero, C., Couzin, I. D., Buhl, J., Kevrekidis, I. G., Maini, P. K., and Sumpter, D.J.T.: Inherent noise can facilitate coherence in collective swarm motion. Proceedings of the National Academy of Sciences 106(14), 5464-5469 (2009)
8. Sun, Y., Lin, W., and Erban R.: Time delay can facilitate coherence in self-driven interacting-particle systems. Physical Review E 90(6), 062708 (2014)

Complex network view of evolving manifolds

Diamantino C. da Silva[1], Ginestra Bianconi[2], Rui A. da Costa[1],
Sergey N. Dorogovtsev[1,3], and José F. F. Mendes[1]

[1] Departamento de Física da Universidade de Aveiro & I3N, Campus Universitário de Santiago,
3810-193 Aveiro, Portugal,
[2] School of Mathematical Sciences, Queen Mary University of London, London, E1 4NS,
United Kingdom
[3] A.F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia

1 Introduction

A convenient digital representation of manifolds (in particular, surfaces) is provided by specific simplicial complexes constructed only of d-dimensional simplexes [$(d+1)$-cliques]. In particular, triangulations provide a convenient digital representation of surfaces. Triangulations are in the very heart of modern civilization providing the main method of treatment of surfaces in topography, engineering, hydrodynamics and aerodynamics, visualization techniques, and everywhere. We use ideas taken from complex networks to generate and describe evolving manifolds by treating triangulations and simplicial complexes as networks with strong constraints (in particular, the conditions that a triangulation network consists only of triangles and that an edge cannot belong more than to two triangles—faces). We observe that these constraints essentially determine the global organization of these networks and influence their local structural properties. In particular, triangulation networks can strongly differ from general planar graphs and networks embedded into two-dimensional metric spaces.

We study complex networks formed by triangulations and higher-dimensional simplicial complexes of closed evolving manifolds [1]. In particular, for triangulations, the set of possible transformations of these networks is restricted by the condition that at each step, all the faces must be triangles. We show that each of these transformations can be performed in a sequence of steps, in which a single elementary transformation is applied in special order. Stochastic application of these operations leads to random networks with different architectures. While previous works were devoted to growing manifolds with boundaries [2–7], here we focus on growing and equilibrium closed manifolds in which every simplex have a chance to be transformed at any instant of the evolution. We perform extensive numerical simulations and explore the geometries of growing and equilibrium complex networks generated by these transformations and their local structural properties. This characterization includes the Hausdorff and spectral dimensions of the resulting networks, their degree distributions, and various structural correlations. Our results reveal a rich zoo of architectures and geometries of these networks, some of which appear to be small worlds while others are finite-dimensional with Hausdorff dimension equal or higher than the original dimensionality of their simplex. The range of spectral dimensions of the evolving triangulations turns out to be from about 1.4 to infinity. Our models include manifolds with evolving topologies, for example, an h-holed torus with progressively growing number of holes. This evolving graph demonstrates features of a small-world network and has a particularly heavy-tailed degree distribution.

2 Models

Our stochastic evolution models are organized in the following way. At each time step, (i) an element or neighboring elements of the simplicial complex under consideration are chosen with some preference or, in the simplest particular case, uniformly at random. For triangulations, such elements are vertices, edges, and triangles. Then, (ii) a specific transformation from the set of operations that keep the simplicial complex intact is applied to this element. For triangulations, this transformation is one of the triangular mesh operations. Depending on specific (i) and (ii), we get a wide range of evolution scenarios, including, in general, growing, decaying, and equilibrium networks with diverse structures, space dimensions, and topologies, that is different, evolving sets of topological features (e.g., a growing number holes in an h-holes torus).

3 Results

Detailed, comprehensive results for all these models are presented in our recent work [1]. Here we focus on model GW from our zoo, which generates a growing triangulation based network homeomorphic to an h-holed torus with progressively growing number of holes. This model is organized as follows. It each step,
(i) choose an edge uniformly at random,
(ii) exchange it for a new vertex attached to all four vertices of the two triangles sharing this edge,
and, in addition, at each θ-th step,
(iii) choose two triangles, excluding first- and second-neighbouring ones, uniformly at random and merge them into a single triangle creating a hole in the manifold (two faces of these triangles annihilate).

The holes in this model play the role of shortcuts in small-world networks, and we observe that the Hausdorff and spectral dimensions of model GW are infinite. The degree distribution of this model decays very slowly, see Fig. 1. In respect of its local properties, this model turns out to be similar to networks whose evolution is driven by aggregation processes [8]. This network has an evolving topology in the sense that new topological features progressively emerge during the evolution.

Other our models provide a wide spectrum of combinations of Hausdorff and spectral dimensions, which can be infinite or finite. Interestingly, we find that our triangulation based networks can have Hausdorff dimensions very different from that for typical planar graphs, which is known to be 4 [9].

We observed that "physical" stochastic network models used for interpretation of evolving simplicial complexes produce a set of surprising results for their local properties (heavy tailed degree distributions) and global ones (unusual values of space dimensions, topological features, holes, coupled with a high local curvature, which is determined by a vertex degree in these networks).

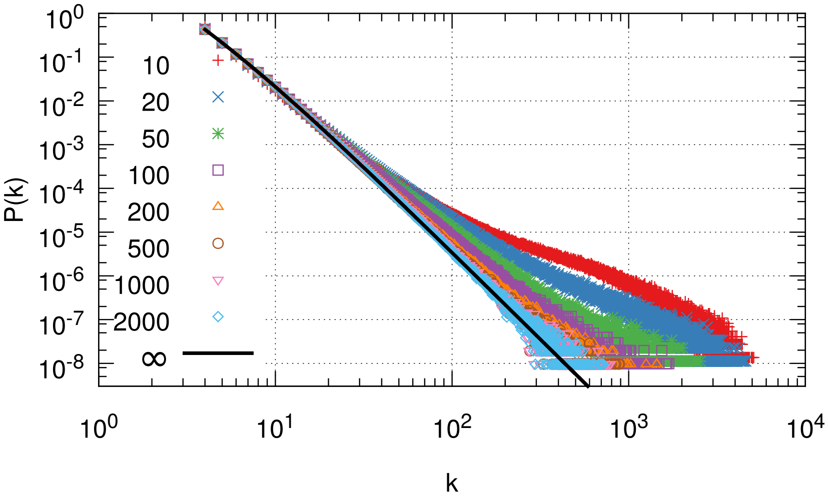


Fig. 1. Degree distribution of the model GW of a growing triangulation network with periodically merging triangles (h-holed torus). Different periods θ for introducing the worm- holes is considered, $\theta = 10, 20, 50, 100, 200, 500, 1000, 2000$, and ∞. The resulting networks in the numerical simulations are of 2^{20} vertices, the averaging is over 10 (for $\theta = 10$) or 100 (for $\theta \geq 20$) samples.

References

1. da Silva, D.C., Bianconi, G., da Costa, R.A., Dorogovtsev, S.N., Mendes, J.F.F.: Complex network view of evolving manifolds. arXiv:1708.02231 (2017)
2. Wu, Z., Menichetti, G., Rahmede, C., Bianconi, G.: Emergent complex network geometry. Sci. Rep. 5, 10073 (2015)
3. Bianconi, G., Rahmede, C., Wu, Z.: Complex quan- tum network geometries: Evolution and phase transitions. Phys. Rev. E 92, 022815 (2015)
4. Bianconi, G., Rahmede, C.: Complex quantum network manifolds in dimension $d > 2$ are scale-free. Sci. Rep. 5, 13979 (2015)
5. Bianconi, G., Rahmede, C.: Emergent hyperbolic geometry of growing simplicial complexes, Scientific Reports 7, 41974 (2017)
6. Bianconi, G., Rahmede, C.: Network geometry with flavor: from complexity to quantum geometry. Phys. Rev. E 93, 032315 (2016)
7. Courtney, O.T., Bianconi, G.: Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. Phys. Rev. E 93, 062311 (2016)
8. Alava, M.J., Dorogovtsev, S.N.: Complex networks created by aggregation. Phys. Rev. E 71, 036107 (2005)
9. Ambjørn, J., Durhuus, B. Quantum Geometry: A Statistical Field Theory Approach. Cambridge University Press, Cambridge (1997)

A branching process with fitness.

Igor E. Smolyarenko

Department of Mathematics
Brunel University
Kingston Lane, Uxbridge UB8 3PH, UK
igor.smolyarenko@brunel.ac.uk,

The model. The model under consideration is the reinforced branching process analysed in a recent series of papers (see [1] and references therein). Variants of this model include the Bianconi-Barabasi (BB) model [2] of preferential attachment with fitness. The model describes a growing population of individuals (e.g., bacteria), each characterised by a fitness $x \in [0,\infty)$ which controls its reproduction rate. Each individual can independently divide with rate x, and each division event can either produce a 'descendant' of the same fitness x (this happens with probability γ), or (with probability β) a 'mutant' with a different fitness y drawn independently from a distribution with density $\mu(y)$. Generically $\alpha+\beta > 1$, so that a division event can produce *both* a descendant and a mutant, with probability $\gamma+\beta-1$. All individuals are immortal. Setting $\gamma = \beta = 1$ reproduces (in continuous time) the dynamics of the degree distribution in the BB model (but not the full topology of the BB network): an 'individual' is now a half-link emanating from a node, with each 'mutant' corresponding to a new node.

A population grown according to these rules can be viewed as a collection of 'families', each with its own fitness (growth rate) x. Quantities of interest include the overall size of the population $N(t)$ at time t, the number of distinct families $M(t)$, the population profile $N(x,t)$, and the distribution of family sizes $P(n,t) = \mathbb{E}\left[\sum_{i=1}^{M(t)} \delta_{n_i(t),n}\right]$, where $n_i(t)$ is the population of the i-th family at time t. An issue that engendered a considerable amount of discussion in the literature is whether this model allows a "winner takes all" behaviour whereby a single family contains a finite fraction of the whole population, analogous to Bose-Einstein (B.-E.) condensation into the lowest available energy eigenstate [2]. This question has been answered in the negative in Ref. [1] in the case of a class of distributions $\mu(x)$ with finite support. The purpose of this note is to investigate the cases of finite and infinite support within a unified framework, explicitly studying the time evolution of the quantities of interest rather than just the limiting behaviour. Expectations of the relevant population sizes are considered, as well as the family size distribution. The distributions of the whole population size require a different set of techniques [3].

Methodology. The starting point is the generating function $G(x,t,z) = \mathbb{E}\left[\sum_{i=1}^{M(t)} z^{n_i(t)}\delta(x-x_i)\right]$, where x_i is the fitness of the i-th family. It follows immediately that $\int dx G(x,t,1) = \mathbb{E}[M(t)]$, similarly $z\partial_z G(x,t,1)|_{z=1} = \mathbb{E}[N(x,t)]$, and $\int dx \oint \frac{G(x,t,z)}{z^{n+1}} \frac{dz}{2\pi i} = P(n,t)$.

Applying the standard techniques one finds that $G(x,t,z)$ satisfies

$$\partial_t G(x,t,z) = \gamma x(z-1)z\partial_z G(x,t,z) + z\beta\mu(x)F(t), \tag{1}$$

with $F(t) = \int \partial_z G(z,x,t)\Big|_{z=1} x\,dx$. Solving Eq. (1) one finds the following self-consistency equation on $F(t)$:

$$F(t) = m(t) + \beta \int_0^t F(\varphi)m(t-\varphi)d\varphi, \tag{2}$$

with $m(t) = \int \mu(x)e^{\gamma xt}xdx$. In order for the population not to explode in finite time $\mu(x)$ must either have a finite support, or decay at large x faster than any exponential.

Case 1: Finite support. Without loss of generality one can restrict x to $[0,1]$. Consequently, $m(t)$ possesses a Laplace transform

$$\tilde{m}(p) = \int_0^1 \frac{x\mu(x)dx}{p-\gamma x}. \tag{3}$$

Equation (2) is then solved straightforwardly:

$$F(t) = \int_{c-i\infty}^{c+i\infty} \frac{dp}{2\pi i} e^{pt} \frac{\tilde{m}(p)}{1-\beta\tilde{m}(p)}, \tag{4}$$

where $c > \max\gamma, p^*$, and the Malthusian parameter p^* (if it exists!) is the root of $1 = \beta\tilde{m}(p)$. Existence of the root distinguishes the regime without "condensation", in analogy with the formalism of genuine B.-E. condensation. It can be shown that p^* exists if $\mu(1)$ is finite (both cases are possible if $\mu(x \to 1) \to 0$), is unique, and that $p^* > \gamma$. This last property ensures that the integral (4) is dominated by the pole at p^* rather than the cut along $[0,\gamma]$, and so $F(t) \to e^{p^*t}/\beta^2\rho(p^*)$, with $\rho(p^*) = \int \frac{x\mu(x)dx}{(p^*-\gamma x)^2}$. Correspondingly, the asymptotically dominant part of the expected population profile evaluates to $\mathbb{E}[N(x,t)] \to \frac{e^{p^*t}}{\beta\rho(p^*)(p^*-\gamma x)}$, with the overall population growth and the total number of families both proportional to e^{p^*t}.

In the opposite case when p^* does not exist ($\mu(x \to 1) \to 0$ is a necessary but not a sufficient condition for that), the integral (4) is controlled by the cut inherited from $\tilde{m}(p)$. The overall population growth is then proportional to $e^{\gamma t}I(t)$, where $I(t) = \int_0^1 \mu(1-\xi)e^{-\xi\gamma t}d\xi$. For example, if $\mu(x) \sim (1-x)^\alpha$ with some finite parameter α, the integral gives a power-law correction to the dominant exponential behaviour. The expected number profile evaluates to the sum of two qualitatively different contributions: a smooth term proportional to $\frac{\beta}{\gamma+\beta}\frac{\mu(x)}{1-x}$, which accounts for a finite fraction of the overall population, and an asymptotically singular term proportional to $\mu(x)e^{-\gamma t(1-x)}$. When t is large (and taking into account that $\mu(x \to 1) \to 0$) the second term has the form of a sharp peak 'squeezed' ever closer to 1. If normalised, this term would converge to a δ-function, consistent with [1], with the peak profile generalising Conjecture 8.1 of [1] obtained for a power-law $\mu(x \to 1)$ in a different setting. The expected number of families under the peak is macroscopic, generalising the conclusion of Ref. [1] that there are no 'winner-take-all' families to arbitary $\mu(x)$ consistent with fintie support and absence of maltusian parameter.

Case 2: Infinite support. The solution of Eq. (2) is now complicated by the fact that Laplace transform of $m(t)$ does not exist. This difficulty can be circumvented by performing an analytical continuation into the complex t plane, solving the equation along a direction where $m(t)$ is 'well-behaved', and analytically continuing the result back to the real axis. (see, e.g., [4]). One therefore obtains $F(t) = \int_{\mathscr{C}_H} \frac{dp}{2\pi i} e^{pt} \frac{\tilde{m}(p)}{1-\beta\tilde{m}(p)}$, where $\mathscr{C}_H$ is the Hankel contour. Subsequent evaluation of the integral gives the following asymptotic result: $\mathbb{E}[N(t)] \to (1+\beta/\gamma)\int_0^\infty \mu(x)e^{\gamma xt}dx$. The asymptotic behaviour of the number profile is subtle, due to the fact that the corresponding normalised density profile does not exhibit uniform convergence as $t \to \infty$. If x is fixed, one obtains $\mathbb{E}[N(x,t)] \to \mathbb{E}[N(t)]\frac{\beta\mu(x)}{\beta+\gamma}$, analogous to the smooth piece in the finite support case. However, the integral of this contribution contains only a finite fraction of the overall population. The second contribution is a smooth peak having the shape $\mu(x)e^{\gamma xt}$, centered at an increasing with time ('travelling') position $x_0(t)$ given by the solution of $\gamma t = -\mu'(x)/\mu(x)$. The area under the peak contributes the 'missing piece' of the overall expected number.

It is illuminating to specialise to $\ln\mu(x) \sim -x^{1+\alpha}$, with $\alpha > 0$ to ensure existence of $m(t)$. One then finds $x_0(t) = (\gamma t/(1+\alpha))^{1/\alpha}$, and, up to subleading corrections, the total number of individuals under the 'travelling' peak is $\exp\{\alpha(\gamma t/(1+\alpha))^{1+1/\alpha}\}$. However, the expected number of families there is proportional to $\exp\{(\alpha-1)(\gamma t/(1+\alpha))^{1+1/\alpha}\}$. Therefore in the regime $1 < \alpha < 2$ this number decays at large t, and eventually becomes less than 1. Thus a family with fitness in the peak region exists only occasionally, with a population much larger than the expected value.

Such a picture may be consistent with occasional existence of a 'winner-take-all' family. Further insight can be gained by analysing the family size distribution:

$$P(n,t) = \int_0^\infty \mu(x)dx \int_{\mathscr{C}_H} \frac{dp}{2\pi i} \frac{\beta F(p)e^{pt}}{\gamma x} \int_{e^{-\gamma xt}}^1 d\zeta\, \zeta^{p/\gamma x}(1-\zeta)^{n-1}. \tag{5}$$

For a finite n this expression reduces to the Beta-function, reproducing the known results [5]. Assuming, however, that n scales with $\mathbb{E}[N(t)]$, one can find that the probability of family size greater than Const. $\cdot\,\mathbb{E}[N(t)]$ behaves as $\exp\{[1-(1+\alpha)^{1+1/\alpha}]\alpha(\gamma t/(1+\alpha)^2)^{1+1/\alpha}\}$, and therefore decays as t becomes large for any α. This decay, however, can be slower than the probability of finding a family under the travelling peak (estimated as the inverse expected number of such families) if $1/(1+\alpha)^{1+1/\alpha} > \alpha$, and therefore not inconsistent with the picture of occasional (with probability going to zero as $t \to \infty$) 'winner-take-all' families, large enough to ensure a finite contribution to the expected overall number. A more subtle analysis exploring the correlations between the numbers of families and their sizes is needed to bring full clarity to the issue.

References

1. S. Dereich, C. Mailler and P. Mörters, *preprint* arXiv:1601.08128.
2. G. Bianconi and A.-L. Barabasi, *Phys. Rev. Lett.* 86, 5632 (2001).
3. I. E. Smolyarenko, *in preparation*.
4. N. M. Temme, *J. Comput. Appl. Math.*, **12** & **13**, 609 (1985).
5. S. Dereich and M. Ortgiese, *Comp. Probab. Comput.* **23**, 386 (2014).

Monitoring polycentric settlement development using a hierarchy-based network indicator

Amin Khiali-Miab[1], Maarten J. van Strien[1], Kay W. Axhausen[2], Adrienne Grêt-Regame[1]

[1] PLUS, ETH Zürich, Stefano-Franscini-Platz 5, 8093 Zürich, Switzerland

[2] IVT, ETH Zürich, Stefano-Franscini-Platz 5, 8093 Zürich, Switzerland

The continuous expansion of settlement areas and the growth of per capita utilization of resources raises serious concerns about the future performance of cities [1, 3, 5]. Meijers [6] shows how different spatial structures of settlement networks have an important impact on the economic performance. Theoretical and empirical studies showed that polycentricty, which refers to the existence of multiple centers in the organization of a system, is associated with the higher efficiency of systems [7], increased territorial cohesion [4] and consequently the higher performance of settlement areas. Polycentricity has become a normative planning goal in many spatial development plans at the global, European and national scales.

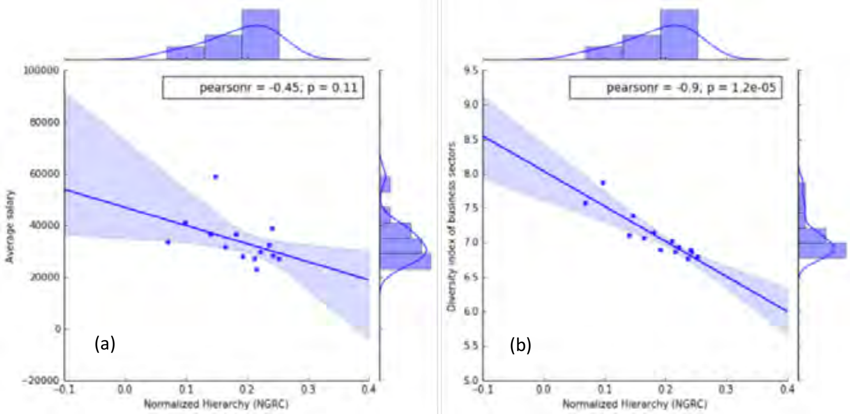


Fig. 1. Correlation between GRC hierarchy measure and (a): average salary, (b): diversity of business sectors

However, spatial planners need quantitative measures to determine whether spatial developments move towards such a polycentric structure or not. This requires polycentricity measures that are robust over different scales and have the capability of predicting the future development of a settlement network. Recent advances in complex network science provide indicators to assess network structures. Hierarchy, for example, is known as the changes of the community structure over different scales and can help to study the distribution of centers in a network. Global Reaching Centrality (GRC) is a hierarchy-based network indicator that has been shown to be capable of differentiating between different types of biological and social networks [8]. In this study, authors demonstrate how GRC can be used to monitor polycentric settlement development. Based on socio-economic data and data from the Swiss national transport model (NPVM), we show (See Fig. 1) how GRC [8], can be used to measure the hierarchical structure of the settlement network in different spatial regions of Switzerland and its relation to the economic activities (specifically salary distribution and business diversity).

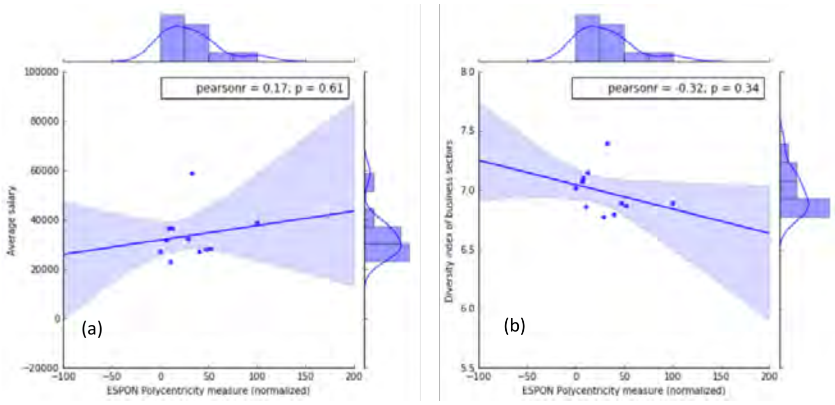


Fig. 2. Correlation between ESPON polycentricity measure and (a): average salary, (b): diversity of business sectors

Comparing GRC and ESPON polycentricity measure [2] (See Fig. 2), GRC is more significantly correlated with the economic activities in a settlement network. Such an indicator might be a first step to support planners assessing spatial urbanization patterns and their relation to economic performance over scale.

References

1. Czamanski, D., Benenson, I., Malkinson, D., Marinov, M., Roth, R., & Wittenberg, L.: Urban sprawl and ecosystems—can nature survive?. International review of environmental and resource economics, 2(4), 321-366 (2008)
2. Dühr, S.: Potentials for polycentric development in Europe: The ESPON 1.1. 1 project report. Planning, Practice & Research, 20(2), 235-239 (2005)
3. Ernstson, H., Leeuw, S. E. V. D., Redman, C. L., Meffert, D. J., Davis, G., Alfsen, C., & Elmqvist, T.: Urban transitions: on urban resilience and human-dominated ecosystems. AMBIO: A Journal of the Human Environment, 39(8), 531-545 (2010)
4. González-González, E., & Nogués, S.: Regional polycentricity: an indicator framework for assessing cohesion impacts of railway infrastructures. European Planning Studies, 24(5), 950-973 (2016)
5. Grêt-Regamey, A., Celio, E., Klein, T. M., & Hayek, U. W.: Understanding ecosystem services trade-offs with interactive procedural modeling for sustainable urban planning. Landscape and Urban Planning, 109(1), 107-116 (2013)
6. Meijers, E.: From central place to network model: theory and evidence of a paradigm change. Tijdschrift voor economische en sociale geografie, 98(2), 245-259 (2007)
7. Meijers, E. J., & Burger, M. J.: Spatial structure and productivity in US metropolitan areas. Environment and planning A, 42(6), 1383-1402 (2010)
8. Mones, E., Vicsek, L., & Vicsek, T.: Hierarchy measure for complex networks. PloS one, 7(3), e33799 (2012)

Part VII

Networks in Finance and Economics

Transfer entropy between communities in complex financial networks

Jan Korbel[1,2,3] and Xiong-Fei Jiang[4,5,6]

[1] Section for Science of Complex Systems, Medical University of Vienna
Spitalgasse 23, 1090 Vienna, Austria
[2] Complexity Science Hub Vienna, Josefstädterstrasse 39, 1090 Vienna, Austria
[3] Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague
Břehová 7, 115 19, Prague, Czech Republic
korbeja2@fjfi.cvut.cz
[4] Department of Physics, Zhejiang University
Hangzhou 310027, P. R. China
[5] School of Information Engineering, Ningbo Dahongying University
Ningbo 315175, P. R. China
[6] Research Center for Finance Computing, Ningbo Dahongying University
Ningbo 315175, P. R. China

1 Introduction

Dynamics of real systems described by large networks, as e.g. financial markets, is a very complex phenomena including non-linear interactions, emergent phenomena and collective behavior. The interactions between the nodes can be measured by several quantities. Among all measures, the most popular are cross-correlations. To the main advantages belong its simplicity and the fact that it describes typical interactions of the system. On the other hand, it lacks of directionality and it can be insensitive to non-linear interactions. This is typically important in situations, when a network becomes into an abnormal regime, as e.g. financial crises in financial markets. Thus, ordinary linear models like correlations cannot reflex the complex nature of the interactions. As a result, a model-free causal measure called *Transfer entropy* has been introduced by Schreiber [1]. It has found many applications in various fields [2, 3]. In many cases, we are particularly interested in the flow of specific parts of the distribution – typically extreme events. For this end, Jizba et al. [4] have introduced *Rényi transfer entropy*, which can accentuate specific parts of information flows.

Complex networks often have an inner structure of strongly interacting nodes. These nodes create communities, and the interaction between communities may be not so significant. On the other hand, inter-community interactions are extremely important in detection of dramatic changes, as extreme falls in financial markets. The community structure can be successfully revealed by PMFG filtering method [6] and InfoMap algorithm [7]. In financial networks as financial markets, it often corresponds to business sectors. Our aim is to investigate transfer entropies between community structures in complex financial networks with focus on information transfer of rare events, which often reflect large movements during abnormal periods, as financial crises. Details of this analysis can be found in Ref. [5].

COMPLEX NETWORKS

The 6^{th} International Conference on Complex Networks & Their Applications. Nov. 29 - Dec. 01, 2017, Lyon (France)

2 Methods

In order to reveal the community structure of the financial market networks, we follow Ref. [8]. As described in introduction, the typical interactions of the network are described by the *correlation matrix* $\mathbb{C}$. Because of the noisy interactions, we make the spectral decomposition $\mathbb{C} = \sum_\alpha \lambda_\alpha \mathbf{u}_\alpha \otimes \mathbf{u}_\alpha$. The largest eigenvalue corresponds to market mode correlation and should be omitted. the smallest eigenvalues correspond to random interaction caused by finite-size effects. The remaining eigenvalues correspond to *sector mode correlations* and are used to obtain the sectors. The community structure is obtained in two steps. First, we filter the full correlation matrix to keep only the most significant interactions. This is done by the PMFG filtering algorithm [6]. The community structure is then estimated by the InfoMap algorithm [7], which is based on random walk in the network.

In order to catch the non-linear nature of interactions, we use the model-free *Transfer entropy*, which is based on Shannon information entropy $S(P) = -\sum_i p_i \ln p_i$. The transfer entropy for discrete time series $Y = \{y(t)\}_{t=1}^N$, $X = \{x(t)\}_{t=1}^N$

$$T_{Y\to X}(m,l) = S(x_{m+1}|y_m,\dots,y_{m-l+1};x_m,\dots,x_1) - S(x_{m+1}|x_m,\dots,x_1) \tag{1}$$

where $S(Y|X) = S(X \cup Y) - H(X)$ is the conditional entropy. Transfer entropy measures information transferred to X uniquely from Y. It can be easily shown that Transfer entropy is always non-negative. When we are particularly interested in information transfer of rare events, we generalize Transfer entropy to *Rényi transfer entropy* introduced by Jizba et al. [4] based on Rényi entropy (RE) $S_q(P) = \frac{1}{1-q}\sum_i p_i^q$. For $q < 1$, RE accentuates tail parts of the distribution, for $q > 1$, RE accentuates central parts of the distributions. Contrary to Shannon transfer entropy, Rényi transfer entropy can be negative, which is equivalent to

$$S_q(x_{m+1}|y_m,\dots,y_{m-l+1};x_m,\dots,x_1) \geq S_q(x_{m+1}|x_m,\dots,x_1) \tag{2}$$

i.e., the extra knowledge of series Y leads to extra risk for the tail parts of the distribution for $q < 1$, or central parts of the distribution for $q > 1$. Negative RTE therefore points to increased complexity caused by emergent non-linear interactions.

3 Results

In Ref. [5], we analyzed five largest financial markets – New York SE, Shanghai SE, Hong Kong SE, Tokyo SE and London SE. For each market we calculate community structure, which corresponds to business sectors. We compare correlations between communities with transfer entropies. As an example, Fig. 1 shows community structure of New York SE and London SE. For each market, we have a correlation structure between communities, Shannon TE and Rényi TE for $q = 0.75$ which accentuates the flow of the rare events [4]. The correlation structure and TE structure remarkably differs. As discussed in [5], the strongest flows are typically observed among financial sectors. Interestingly, the largest flows between communities exhibit the most negative RTE, which points to the fact that the rare events play the major role in information transfer based on TE (contrary to correlations).

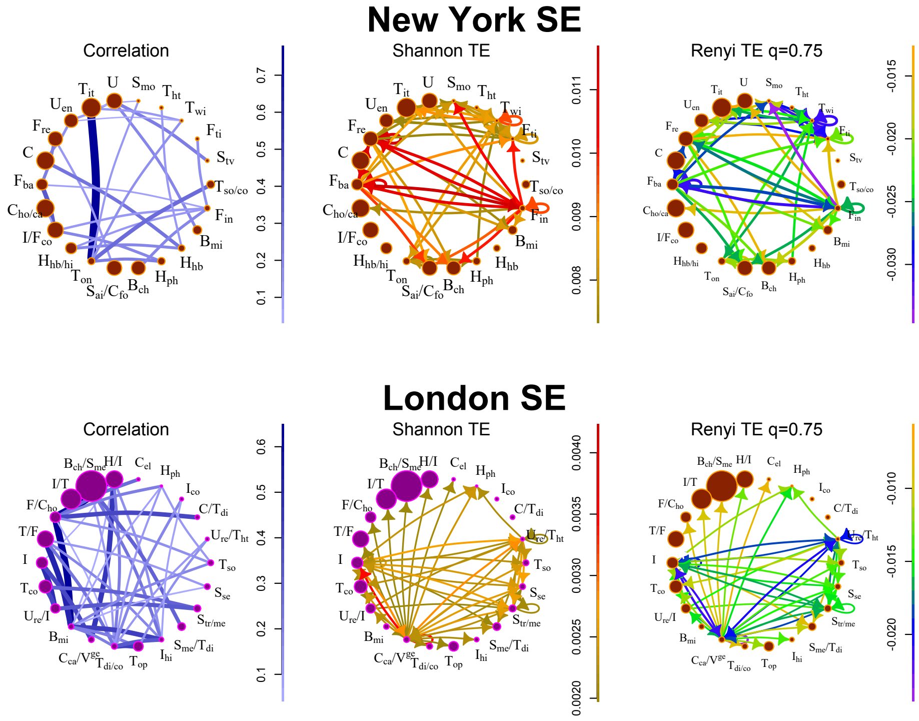


Fig. 1. Community structure of New York SE and London SE and comparison of correlation and transfer entropy networks for the communities.

Acknowledgements

J. K. was supported by the Austrian Science Fund, grant No. I 3073-N32 and the Czech Science Foundation, grant No. 17-33812L. X.-F. J. was supported by NNSF of China, grant No. 11505099.

References

1. T. Schreiber. *Phys. Rev. Lett.* **85** (2000), 461.
2. R. Marschinski and H. Kantz. *Eur. Phys. J. B* **30** (2002),275.
3. K. Hlaváčková-Schindler, M. Paluš, and M. Vejmelka. *Phys. Rep.* **441(1)** (2007), 1.
4. P. Jizba, H. Kleinert, and M. Shefaat. *Physica A* **391(10)** (2012), 2971.
5. J. Korbel, X.-F. Jiang and B. Zheng, *arXiv:1706.05543*.
6. M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. *PNAS* **102** (2005), 10521.
7. M. Rosvall and C. T. Bergstrom. *PNAS* **105** (2008), 1118.
8. X. F. Jiang, T. T. Chen, and B. Zheng. *Sci. Rep.* **4** (2014), 5321.

Business cycles' correlation in the Japanese production network

Hazem Krichene[1], Abhijit Chakraborty[1] Hiroyasu Inoue[1], and Yoshi Fujiwara[1]

Graduate School of Simulation Studies, University of Hyogo, Japan
krichene.hazem@gmail.com

1 Introduction

Several recent works showed the importance of network structure in the modeling of economic patterns. [3] showed that the general business cycle dynamics depend on the microscopic properties of the economic system. Thus, business cycle fluctuations and shocks propagation can be reflected by the production network of the economy. Such production network for the Japanese economy was considered in several works as in [2]. These economic shocks propagation were studied by [1, 4] who showed that the Japanese business cycle are correlated between prefectures, based on their GDP levels' similarities. In this work we propose to combine these two literature fields to study the risk of dependencies in the Japanese economy through a new view. We consider that business cycle correlations are related to the community structure of the production network. Communities are defined based on the Infomap algorithm (see [6]). Then, communities are specified by their respective GDPs, which are approximated by the sum of total sales of their firms. By applying a Hodrick-Prescott filter, the business cycles of all communities are captured, and their correlations are modeled based on their bivariate joint distributions using copula theory. Copula theory is used to capture all the dependency structures of business cycles due to the non-normality of GDP fluctuations, as found in some empirical works providing stylized facts about GDP. Finally, the inter- and intra-community business cycle correlations are explained by different linear econometric models.

2 The production network data

The data for the Japanese production network are based on a 2016 survey by Tokyo Shoko Research (TSR). A link from firm i to firm j means that i is the supplier of j, and j is the customer of i. We note that the links are unweighted. To construct the GDP of the production network, the total sales amounts are needed which are available on the Profit-Loss statement of each firm. The Nikkei Digital Media database is used, which contains PL statements for firms listed on the Tokyo Stock Exchange from 1980 to 2012. Only active firms in that period are considered which give a sub-production network of 940 firms with 5,431 links. The production network shows a scale-free behavior and a disassortative mixing. The constructed business cycle shows a high correlation with the actual Japanese business cycle observed between 1980 and 2012 (the regression gives a coefficient of 1).

3 Results and discussion

The Infomap algorithm reveals 73 communities with 1,028 inter-communities links. Their correlations are given by the Kendall tau estimated based on their best bivariate copula. Fig. 1 compares the inter- (denoted by communities α, β) and intra-community (denoted by firms i, j) correlations. The inter- and intra-community correlations are very close showing the significant dependencies at the micro and macro levels of the economy.

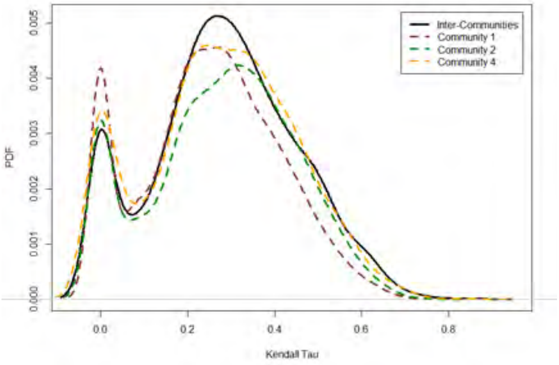


Fig. 1. Probability density functions of the business cycle correlations in the Japanese production network.

The determinants of these business cycle correlations are studied based on economic and topological variables through several linear econometric models. At the inter-communities level, we consider four economic variables to explain their business cycle correlations: sector and geographic homophily based on the Jensen-Shannon distance (see [5]) and GDP and community size (number of firms per community) heterophily based on a weighted absolute value of the difference. At the intra-community level, three topological variables are considered: firms degree and clustering heterophily, and shortest path. Results are exposed in Table 2. We found that the inter-community business cycle correlations increase with sector and geographic similarities and with GDP heterophily. In fact, due to the disassortative mixing of the network, small communities are linked to large communities, which increases their dependence and their potential vulnerability in case of disastrous crisis. In the other side, the intra-community business cycle correlations were studied through four selected communities (the largest communities, i.e. other communities are too small to be considered for estimation with linear econometric model). For the largest communities (community 1 and community 2) in the network, we showed significant positive impacts of the shortest distance and clustering similarity on business cycle correlations. Moreover, the impact of disassortative mixing was confirmed at the firm level, where correlations increase with the firm degrees heterophily. However, non-significant explanation are found for the business cycle correlations of the small communities (community 4 and community 6).

Table 1. The first two columns concern the inter-communities correlation, while the last five columns concern the intra-communities correlation. *, ** and *** indicate significant values at the 1%, 5% and 10% levels

economic variables	$\tau_{\alpha,\beta}$	topological variables	τ^1_{ij}	τ^2_{ij}	τ^4_{ij}	τ^6_{ij}
$\text{JS}^{sector}_{\alpha,\beta}$	-0.29^{***} (0.03)	$\text{Degree}^{heterophily}_{ij}$	0.07^{***} (0.01)	0.02^{*} (0.01)	0.01 (0.02)	0.09 (0.08)
$\text{JS}^{geo}_{\alpha,\beta}$	-0.17^{***} (0.02)	Path_{ij}	-0.06^{***} (0.01)	-0.22^{***} (0.02)	-0.03 (0.03)	-0.15 (0.11)
$D^{\text{GDP}}_{\alpha,\beta}$	0.13^{**} (0.04)	$\text{Clustering}^{heterophily}_{ij}$	-0.02^{**} (0.01)	-0.04^{**} (0.01)	-0.04^{**} (0.02)	0.03 (0.03)
$D^{S}_{\alpha,\beta}$	-0.004 (0.04)	intercept	0.29^{***} (0.01)	0.38^{***} (0.01)	0.31^{***} (0.02)	0.28^{***} (0.01)
intercept	0.58^{***} (0.03)	-	-	-	-	-

4 Conclusion

Understanding how different economic agents, such as banks and firms, are correlated is of high importance to measuring the systemic risk of a country. In a new contribution, we considered the Japanese supplier-customer network to classify groups based on the community structure determined by the Infomap algorithm. These communities reflect the proper group-based structure of the Japanese production network, which is based on both the sector and geography. The results showed evidence of significant business cycle correlations at the inter- and intra-community levels. These correlations were very similar, indicating that dependence risks are significant between firms in the same community and spread to other communities. Finally, these correlations were explained based on economic and topological variables and we found that the significant systemic risk in the Japanese economy is more likely to be disastrous for small communities and small firms.

Table 2. Estimation results of ERG model applied on production network of firms belong to the Tokyo Stock Exchange. All endogenous and exogenous attributes are highly significant based on the p-value.

Attributes	$\hat{\theta}_{MLE}$	$s.e(\hat{\theta}_{MLE})$	p-value	Attributes	$\hat{\theta}_{MLE}$	$s.e(\hat{\theta}_{MLE})$	p-value
Degree	-4.05	0.01	≤ 0.01	Profit Sender	$2.31 \cdot 10^{-07}$	$9.61 \cdot 10^{-10}$	≤ 0.01
Reciprocity	1.15	0.004	≤ 0.01	Profit Receiver	$1.02 \cdot 10^{-07}$	$4.10 \cdot 10^{-10}$	≤ 0.01
In-Stars	2.01	0.004	≤ 0.01	Out-Stars	2.04	0.005	≤ 0.01
Sector Homophily	0.59	0.005	≤ 0.01	Location Homophily	0.16	0.002	≤ 0.01
Two-Path	-0.02	$9.08 \cdot 10^{-05}$	≤ 0.01	Profit Heterophily	$8.37 \cdot 10^{-08}$	$4.43 \cdot 10^{-10}$	≤ 0.01
AT-T	0.06	0.0007	≤ 0.01	AT-U	0.17	0.0005	≤ 0.01
AT-C	0.13	0.002	≤ 0.01	AT-D	0.20	0.0008	≤ 0.01

References

1. Artis, M., Okubo, T.: The intranational business cycle in Japan. Research Institute for Economics and Business Administration, Kobe University. Technical Report, (2009)
2. Fujiwara, Y., Aoyama, H.: Large-scale structure of a nation-wide production network. The European Physical Journal B, 77(4):565580 (2010)
3. Gabaix, X.: The granilar origins of aggregate fluctuations. Econometrica, 79(3):733–772 (2011)
4. Ikeda, Y., Aoyama, H., Iyetomi, H., Yoshikawa, H.: Direct evidence for synchronization in Japanese business cycles. Evolutionary and Institutional Economics Review, 10(2):315–327 (2013)
5. Lin, J.: Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory, 37(1):145–151 (1991)
6. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences,105(4):1118–1123 (2008)

Factor modeling of financial asset returns for partial correlation network

Takashi Isogai[1]

Bank of Japan, 2-1-1 Hongoku-cho Nihonbashi Chuo-ku Tokyo, Japan
takashi.isogai@gmail.com

1 Introduction

Correlation network analysis of financial asset returns including stock prices and exchange rates is a very useful tool to clarify the underlying structure of those financial markets ([1]). The choice of correlation measure, namely, correlation or partial correlation, depends on the purpose of analysis of individual relationship. The partial correlation is employed when we are interested in the degree of association between two assets excluding any indirect channel of comovements due to common driving factors. The correlation network of financial asset returns tends to be a dense network because of a very high level of correlation observed market-wide. The network may become more sparse when converted to a partial correlation based one; the local correlation structure would be more apparent. There is, however, a difficult problem to calculate the partial correlation between asset returns, since those common factors cannot be observed directly. Thus, we need to develop some factor model of asset returns to control for the systematic contribution of common factors to calculate the partial correlation. We propose a new factor modeling framework based on network clustering results achieved in our previous research ([2]). The method enables to build a partial correlation network by decomposing asset returns into systematic component and unsystematic (idiosyncratic) component.

2 Factor modeling for partial correlation

In the context of financial modeling, a conventional factor modeling that uses a market asset price index that is calculated as a simple (or some weighted) mean of all asset returns as the market-wide common factor. A regression type model is used to remove the systematic factor contribution from the correlation measure to calculate the partial correlation. This method is easy to be implemented, but is not flexible to control for more complicated common driving factors. If we know something about market segmentation, we can identify such common movements of asset returns in a more accurate and flexible way. In our previous research, we developed a data-driven framework to achieve correlation network clustering of Japanese stock returns, which can provide such market segmentation information ([2]). Unlike the existing standard industrial sector segmentation that is not necessarily reliable classification of stocks, the proposed method works well to find highly correlated stock groups. The method employs a hierarchical network division which utilizes modularity as the divisive criteria as illustrated by Fig. 2: the whole Japanese stock market is divided into 14 groups (marked as A-N). In this study, we use this stock group information to develop factor modeling of stock returns using the Japanese stock return data, while the method can be applied to other financial assets.

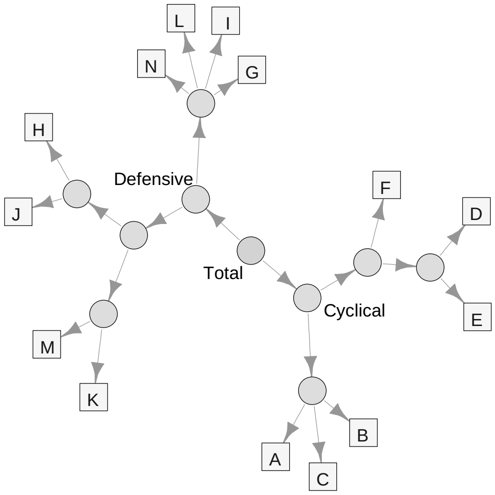


Fig. 1. Clustering by recursive network division (Japanese stock market)

The classification result is used as supervising information for dimensionality reduction of factor modeling. Specifically, we apply the supervised principal component analysis (SPCA) to the filtered (centered and standardized by using a financial volatility model) stock returns ([3], [4]). It is possible to apply PCA to the whole of stock returns in the market to extract global common factors. In this case, the same common factors affect all stocks in any group. In contrast, our method applies SPCA to every stock group independently to extract common factors. The group information is used to eliminate the factor contribution to dependent variable (returns) from other groups. Those common factors are, thus, defined as some mixture of a market-wide global factor and local group-wide factors that vary depending on the groups. The SPCA is represented in the form of singular value decomposition (SVD) as

$$X_\theta = U_\theta \Sigma_\theta V_\theta^T \tag{1}$$

where X_θ is the return series data matrix that has columns as return time series, $U_\theta \Sigma_\theta$ is the principal components (scores) regarded as the common driving factors, V_θ is the factor loading. θ identifies stock groups marked as from A to N in Fig. 2. Individual returns in each group are then fitted to a linear regression model with dependent variables of some of those principal components; thus, returns are separated into systematic components and residuals. The factor model is represented as

$$z_{i,\theta} = C_\theta \ \beta_{i,\theta} + \varepsilon_{i,\theta} \tag{2}$$

where i is stock id in group θ, C_θ (explanatory variables in a reduced dimension) is a subset of $U_\theta \Sigma_\theta$ in (1), $\beta_{i,\theta}$ is regression coefficients, and $\varepsilon_{i,\theta}$ is residuals (idiosyncratic factor).

The partial correlation between returns is calculated based on the residuals $\varepsilon_{i,\theta}$ in (2) to calculate a partial correlation network adjacency matrix. The model is flexible in that a broader (narrower) group definition can be used as θ. The definition of systematic factor is rather contingent on an analytical scope; therefore, flexible coverage of θ could help establish flexible factor modeling. For example, returns of a stock in group A can be modeled using a broader parent group as θ that includes B and C as well as A. Thresholding the number of principal components of $U_\theta \Sigma_\theta$ to setup C_θ is another important issue related to dimensionality reduction.

Once residuals are separated from returns by eliminating common factors contribution, sparse partial correlation networks are identified, which represent more direct linkages between stocks. A more detailed analysis on topological structure and local network community detection can be conducted based on the partial correlation network adjacency matrix. We apply the above method to more than 1,300 Japanese stock returns to build partial correlation networks. The partial correlation networks are compared with the one built using the simple global factor model to clarify differences between the two approaches.

3 Conclusion

The separation between systematic and idiosyncratic components of financial asset return is an important step to build a partial correlation network of returns. A simple one common factor modeling is widely used there; however, a more flexible method of factor modeling required to build a sparse but structured correlation network. Our proposed method based on SPCA is useful for that purpose. This research is still at an early stage; therefore, the method should be tested more widely on the real data. The usage of partial correlation network of financial asset returns should be discussed more in the context of factor modeling, which can be applied to portfolio optimization and risk control.

References

1. Kenett, DY., Huang, X., Vodenska, I., Havlin, S., Stanley, HE.: Partial correlation analysis: Applications for financial markets. Quant. Finan. 15(4), 569578 (2015)
2. Isogai, T.: Clustering of Japanese stock returns by recursive modularity optimization for efficient portfolio diversification. J. Complex. Netw. 2(4), 557-584 (2014)
3. Bair, E., Tibshirani, R.: Semi-supervised methods to predict patient survival from gene expression data. PLoS biology, 2(4), e108. (2004)
4. Bair, E., Hastie, T., Paul, D., Tibshirani, R.: Prediction by supervised principal components. J. Am. Stat. Assoc. 101(473), 119–137 (2006)

Irresponsible communities in global supply chain

Takayuki Mizuno[1], Takaaki Ohnishi[2], and Tsutomu Watanabe[3]

[1] National Institute of Informatics, Tokyo 101-8430, Japan,
mizuno@nii.ac.jp
[2] Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan
[3] Graduate School of Economics, The University of Tokyo, Tokyo 113-0033, Japan

1 Introduction

What are the responsibilities of multinational corporations amid economic globalization in dealing with environmental degradation, widespread poverty, refugee displacement, racial discrimination, terrorism, regional conflicts, and other global challenges? Certainly one initiative that would nudge the world in the right direction would be for companies to assume greater responsibility for the environment and human rights issues by improving the transparency of entire supply chains from a global perspective. For example, United Kingdom wants to prohibit the purchase of materials from irresponsible companies involved with regions that exploit slave labor or otherwise disregard the human rights of workers [1]. Bolstering transparency contributes to the production of high-quality products, improves long-term business efficiency, and helps establish sustainable supply chains [2,3]. More than 90% of the world's companies belong to a single global supply chain, and any two companies selected at random from this supply chain are only separated by about six customers on average [4,5]. Since the global supply chain has this feature of small-world connectedness, it is exceedingly difficult to check every supplier and client upstream from each publicly traded UK firm. We propose a scheme that markedly improves transparency by exploiting this network structure.

In this paper, we merge the global supply chain dataset provided by S&P Capital IQ with a negative information dataset provide by Dow Jones about 50,000 irresponsible companies that are widely covered by the global media-such as media reports of companies employing child labor. We discuss situations in which an irresponsible company is established in a particular community making up the global supply chain. Identifying the companies that bridge this community and industries in developed countries, and stopping the irresponsible flow through these companies, are key to solving the issue.

2 Results

Global supply chain (=global customer-supplier network) is built by multiple communities. We use the map equation with multilevel algorithm to detect communities in this network that covers about 400,000 major incorporated firms, including all the listed firms in the world. The global supply chain can be divided into communities up to four layers. However, the layer for most communities is three. Figure 1 shows the community size distributions on the first layer, and on the second layer in the largest community on the first layer. The community size on the first layer follows the truncated power law distribution. In the top 20 communities of size on each layer, 63% of all firms on the first layer and 13% of all on the second layer are included. We statistically investigate the attributes (industry type / address) of firms included in each community on each layer. On the first layer, the firms comprise a community with those firms that belong to the same industry but different home countries, indicating the globalization of firms' production activities. On the other hand, communities representing countries and regions are often observed on the second and third layers.

Next, we investigate how many irresponsible companies that are widely covered by the global media are included in each community. If irresponsible companies are uniformly distributed in the global supply chain, the number of irresponsible companies in each community follows the Poisson distribution. We count the irresponsible companies listed in the data set provided by Dow Jones in community i on the lowest layer, and set it to n_i. We measure the degree of irresponsibility d_i of community i as follows,

$$d_i = (n_i - ps_i)/\sqrt{ps_i} \tag{1}$$

where s_i is size of community i and p indicates "(the number of irresponsible companies)/(total number of companies)". Figure 2 shows the distribution of degree of irresponsibility d_i. The red line represents the Poisson distribution which normalized by the mean and the standard deviation. The d_i is negative in many communities. That is, many communities contain few irresponsible companies. On the other hand, there are small number of communities including many irresponsible companies that exceed 3σ and 5σ of the normalized Poisson distribution. These results mean irresponsible companies are consolidated in specific communities. It is important to identify the companies that bridge these communities and industries in developed countries, in order to prevent inflow of products made by irresponsible companies into the developed countries.

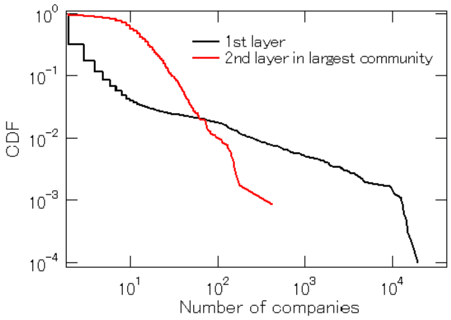


Fig. 1. Community size distributions on the first layer, and on the second layer in the largest community on the first layer in global supply chain.

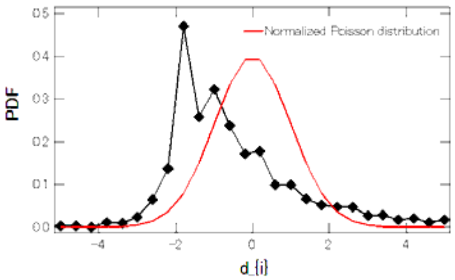


Fig. 2. Irresponsibility degree distribution of communities. The red line represents the Poisson distribution which normalized by the mean and the standard deviation.

References

1. Modern Slavery Act 2015 in the United Kingdom.
2. Hannah Koep-Andrieu, OECD Global Forum on Responsible Business Conduct 2017.
3. Samir K. Srivastava, Green Supply Chain Management: A state-of-the-art Literature Review, International Journal of Management Reviews 9(1), 53-80, 2007.
4. Takayuki Mizuno, Takaaki Ohnishi and Tsutomu Watanabe, Structure of global buyer-supplier networks and its implications for conflict minerals regulations, EPJ Data Science 5, 2 (15 pages), 2016.
5. Takayuki Mizuno, Takaaki Ohnishi, Tsutomu Watanabe, The Structure of Global Inter-firm Networks, Social Informatics Lecture Notes in Computer Science 8852, pp.334-338, 2015.

International Trade Relationship from a Multilateral Point of View

Hiroshi Iyetomi[1], Yuichi Ikeda[2], Takayuki Mizuno[3],
Takaaki Ohnishi[4], and Tsutomu Watanabe[5]

[1] Department of Mathematics, Niigata University, Niigata 950-2181, Japan,
hiyetomi@sc.niigata-u.ac.jp
[2] Graduate School of Advanced Integrated Studies in Human Survivability, Kyoto University,
Kyoto 606-8306, Japan
[3] National Institute of Informatics, Tokyo 101-8430, Japan
[4] Graduate School of Information Science and Technology, The University of Tokyo,
Tokyo 113-8656, Japan
[5] Graduate School of Economics, The University of Tokyo, Tokyo 113-0033, Japan

1 Introduction

Countries are connected through their trade relations, forming a bidirected network in which nodes can be linked in both ways. The progressive globalization of the world economy makes the trade relationship more and more complex. Structural properties of the world trade network and its evolution have thus attracted interest of researchers in network science (see [2] and references therein).

Imbalance of bidirectional trade flows linking two countries may cause trade friction between them. However, reciprocity between the two countries may emerge if one regard the bilateral relation as a part of the complex trade network. Here we like to demonstrate empirically how much such a multilateral point of view change our understanding of the international trade relationship. As will be appreciated, a network-theoretic approach plays a central role in this study.

2 Helmholtz-Hodge Decomposition

We apply the Helmholtz-Hodge decomposition (HHD) [4] of flow on a network to the trade imbalance network; it uniquely breaks up the flow into potential and loop components. The potential flow elucidates hierarchical structure in the global trade. On the other hand, the loop flow quantifies the degree of reciprocity among countries.

Figure 1 illustrates how the HHD works with a triangular trade relationship among three countries as shown by the diagram on its left-hand side; country A imports goods valued at 3 in given monetary units from country B and so forth. The two diagrams on the right-hand side are outcomes of the HHD applied to the triangular trade. The first diagram, in which each country (node) is accompanied by its potential value ϕ, shows the potential flow component; the flow between two countries is given by difference between the potentials of the two countries. The potential values thus enable us to align the three countries in the order A→B→C from upstream to downstream. The second diagram depicts the loop flow component, in which the surplus due to outgoing flow is exactly compensated by incoming flow at each node.

If we focus on the bilateral relationship, country A has a trade deficit against country
B. Howe he trade
conflict l A even
generate

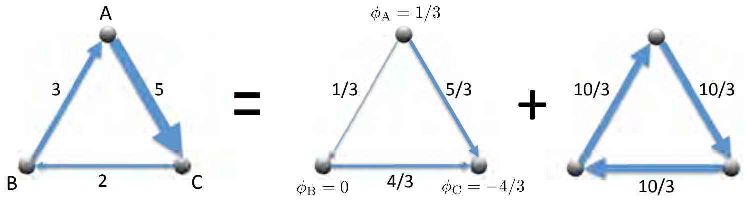


Fig. 1. Application of the Helmholtz-Hodge decomposition to a triangular trade network.

3 Results and Discussion

We use the trade data compiled by Gleditsch [3], who extended the DOTS data set of IMF [1] with alternative data sources to make his data as comprehensive as possible. Unfortunately, Gleditsch's data set cover years up to 2000 starting from 1948. To construct the trade networks after 2000, we trace back to the original DOTS data set.[6]

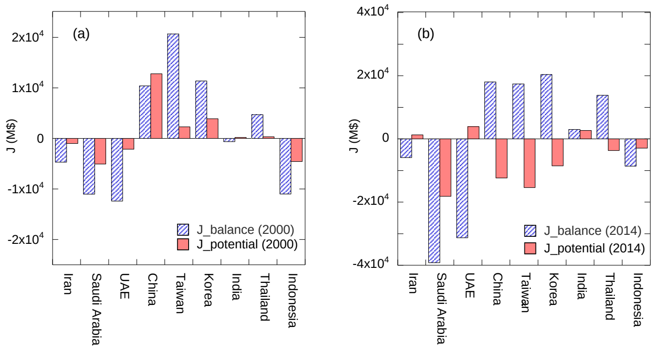


Fig. 2. Trade relationship of Japan with Asian countries in 2000 (a) and 2014 (b). The visible trade balance (blue bar) is compared with the effective trade balance (red bar) determined by the Helmholtz-Hodge decomposition. A positive (negative) value means that Japan has a trade balance surplus (deficit) against its counterpart.

Figure 2 shows the HHD significantly reduces the imbalance in the trade relations between Japan and Asian countries except for China in 2000. The multilateral point

[6] We separately collected the trade data of Taiwan, which is a nonmember of IMF.

of view changes the trade situation of Japan with Asian countries more drastically in 2014. Directly, Japan generated large surpluses from trade with China, Taiwan, and Korea. In effect, however, Japan had notable trade deficits with those countries caused by additional indirect paths through other countries. Also Fig. 3 shows trade relations among the top 33 countries with respect to GDP in 2014. The countries are aligned in the vertical direction according to their potential values obtained by the HHD for the trade imbalance network; countries positioned at the upper (lower) side are regarded relatively as exporters (importers). Japan is positioned lower than the three Asian countries in accordance with the results in Fig. 2. The positions of countries in the horizontal direction reflect strength of the trade relations among them; more tightly coupled are two countries through trade, more closely located are they.

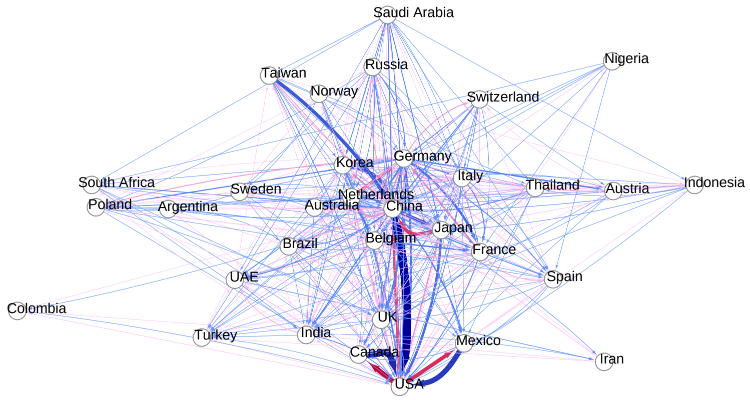


Fig. 3. Trade relations among the top 33 countries with respect to GDP in 2014. Blue arrow depicts trade flow from a upper country to a lower county in the Helmholtz-Hodge ranking, and red arrow shows trade flow in the other way around. The width of each arrow is proportional to the corresponding trade flow; trade flows smaller than 1% of the largest flow are not shown.

The HHD thus enables us to delve into the global trade relationship with a multilateral perspective; the loop flow component in the trade imbalance network is a possible clue to "mystery of the excess trade balances" in international macroeconomics.

We thank Dr. Yuichi Kichikawa for his help in preparing Fig. 3. This research was supported by a grant-in-aid from Zengin Foundation for Studies on Economics and Finance and by JSPS KAKENHI Grant Number 16K13375.

References

1. Direction of Trade Statistics (International Monetary Fund): `http://data.imf.org/?sk=9D6028D4-F14A-464C-A2F2-59B2CD424B85`
2. Duenas, M., Fagiolo, G.: Global trade imbalances: A network approach. Advances in Complex Systems (ACS) 17(03n04), 1–29 (2014)
3. Gleditsch, K.S.: `http://privatewww.essex.ac.uk/~ksg/exptradegdp.html`
4. Jiang, X., Lim, L.H., Yao, Y., Ye, Y.: Statistical ranking and combinatorial hodge theory. Mathematical Programming 127(1), 203–244 (2011)

Countries' positions in the international global value networks: Centrality and economic performance

Isabella Cingolani[1], Pietro Panzarasa[2], and Lucia Tajoli[3]

[1] Big Data and Analytical Unit, Imperial College London,
[2] School of Business and Management, Queen Mary University of London,
[3] Department of Management, Economics and Industrial Engineering, Politecnico di Milano

1 Introduction

The increasing relevance of global value chains (GVCs) in international trade has been widely emphasised by a number of recent studies (e.g.,[1]). GVCs are the result of production processes stretching across multiple countries, such that different phases of the production can exploit the specific comparative advantages of each location. In this work, we draw on data concerned with world trade in three different sectors to uncover the network structure and dynamics of GVCs. Our aim is to assess the centrality of countries in the global network of links generated by international production processes.

While centrality measures have mostly been developed for one-mode, binary, and time-invariant networks, there have been a number of attempts to extend such measures to two-mode networks [2], or more generally K-partite graphs, as well as weighted [3], multiplex ([4]), and time-varying networks [5]. Attempts have also been made to cross-classify centrality measures in terms of various criteria, such as their underlying assumptions on how processes unfold in a network [6]. Recently, a number of studies have drawn upon the network literature to develop suitable sets of metrics to capture the role that countries play in the international GVCs [7, 8]. Yet, what is still largely missing is an analytical framework in which the centrality of a country is explicitly formalised as a multi-faceted measure that captures the country's position at the various production stages into which the international GVCs are organised.

2 Formalising the centrality of countries in the international global value networks

A necessary step towards the assessment of the centrality of countries in GVCs is represented by the extraction of the global value networks (GVNs) from the underlying international trade networks. To this end, we define the tripartite valued graph whose nodes are partitioned into three different independent sets, U, M, and D. The first set U refers to the population of exporters of intermediate inputs. The second set M refers to the population of countries that are importers of intermediate products or exporters of finished products or both importers of intermediate products and exporters of finished products. Finally, the third set D includes the importers of finished products. The roles that countries occupy within the international GVNs can be unmasked through the application of suitable centrality measures to the international trade network of intermediate and finished products.

Here we propose three measures of centrality applied to the directed and weighted international global value tripartite network defined above for capturing the degree to which a given country plays a prevailing role in the upstream, midstream, and downstream stages of the production in a given industry's GVN: (a) a country's *upstreamness* centrality in an industry captures the tendency of the country to preferentially export intermediate goods to other countries that, in turn, have a tendency to preferentially export finished products and import intermediate inputs; (b) a country's *downstreamness* centrality in an industry captures the tendency of the country to preferentially import final products from countries that, in turn, tend to preferentially export finished products and import intermediate inputs; and (c) a country's *midstreamness* centrality in an industry captures the tendency of the country to import intermediate goods preferentially from countries with high upstreamness centrality and to export final products preferentially to countries with high downstreamness centrality.

3 Results

Our study draws upon the bilateral trade data set extracted from the BACI-CEPII database. We restricted our analysis to the trade flows in three industrial sectors: Electronics, Motor Vehicles, and Textiles and Apparel. We distinguished between "intermediate goods" and "finished products", and used these two categories to extract the international GVNs from the international trade network. In the case of Electronics, Fig. 1 suggests that China is the second-ranked country in terms of imports, but it is not equally central as a final market since it holds a much lower-ranked position in downstream centrality. Somewhat similar profiles are the ones of Taiwan and Malaysia, countries that import many electronic intermediate products but are not equally relevant as a final market. On the contrary, countries such as Mexico and Brazil are more central in the intermediate positions of the international production network of electronic goods than their trade volumes would lead us to expect, while many advanced European countries are much less central. In addition, our results suggest that the correlation between more traditional measures of trade, such as exports and imports, and our centrality measures is positive, but far from perfect.

4 Conclusions

To fully assess a country's position in the international production of goods and services, it is crucial to evaluate the country's centrality at the various stages into which the GVCs can be articulated. To this end, we proposed a novel three-faceted measure of centrality that captures a country's distinct roles at the upstream, midstream, and downstream stages of the international production process. Our findings suggest that countries hold different positions at the various stages of the international production process, and these positions change over time. These variations in centrality according to the roles countries occupy along the GVNs would remain undetected if more traditional measures of market power based on aggregate trade values were used.

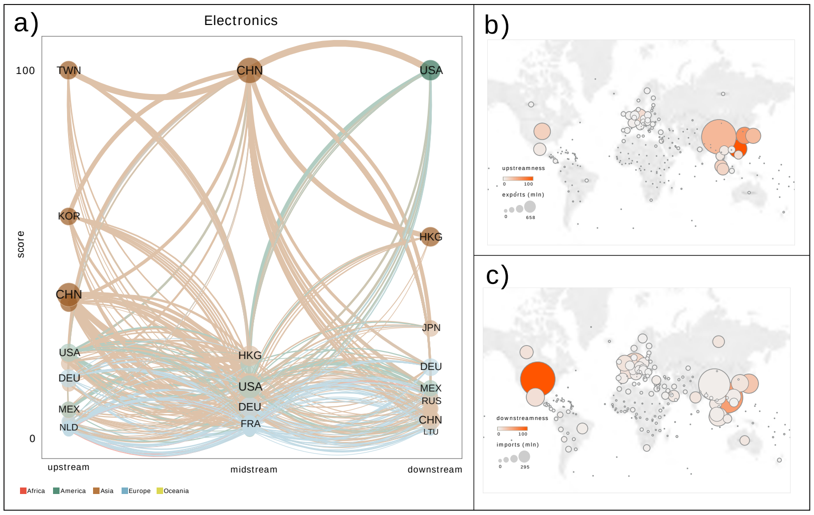


Fig. 1: Rankings and values of upstreamness centrality, midstreamness centrality, downstreamness centrality, exports, and imports of countries in the Electronics industry in 2014. In panel (a), the size of each node in the network is proportional to the sum of the country's exports of intermediate inputs (upstream), the sum of the country's imports of intermediate inputs and exports of final products (midstream), and the sum of the country's imports of final products (downstream). For each centrality measure, countries are ranked according to the corresponding score (highest at the top). The width of each link is proportional to the value of products exchanged by the two connected countries. The colour of each link refers to the continent of the country from which the link originates. Panels (b) and (c) show the geographic map in which each country is represented as a circle whose diameter is proportional to the country's total exports (b) and total imports (c), and whose colour varies according to the corresponding value of upstreamness (b) and downstreamness (c) centralities.

References

1. Baldwin, R, Lopez-Gonzalez, J (2015) Supply-chain trade: A portrait of global patterns and several testable hypotheses. *World Econ* 38(11): 16821721.
2. Faust, K (1997) Centrality in affiliation networks. *Social Networks* 19: 157-191.
3. Barrat, A, Barthélemy, M, Pastor-Satorras, R, Vespignani, A (2004) The architecture of complex weighted networks. *Proc Natl Acad Sci U S A* 101(11): 3747-3752.
4. Boccaletti, S, Bianconi, G, Criado, R, del Genio, C, Gomez-Gardeñes, J, Romance, M, Sendina-Nadal, I, Wang, Z, Zanin, M (2014) Structure and dynamics of multilayer networks. *Phys Rep* 544: 1.
5. Nicosia, V, Tang, J, Mascolo, C, Musolesi, M, Russo, G, Latora, V (2013) Graph metrics for temporal networks. In: Holme, P, Saramäki, J (eds) *Temporal Networks. Understanding Complex Systems*, 15-40. Springer-Verlag, Berlin Heidelberg.
6. Borgatti, SP (2005) Centrality and network flow. *Social Networks* 27: 55-71.
7. Koopman, R, Wang, Z, Wei, SJ (2014) Tracing value-added and double counting in gross exports. *Am Econ Rev* 104(2): 459-494.
8. Lejour, A, Rojas-Romagosa, H, Veenendaal, P (2014) Identifying hubs and spokes in global supply chains using redirected trade in value added. Working Paper Series 1670, European Central Bank.

Part VIII

Motif Discovery and Link Analysis

Entropy-based approach to missing links imputation

Federica Parisi[1], Guido Caldarelli[1], and Tiziano Squartini[1]

IMT School for Advanced Studies, Lucca, Italy

1 Introduction

A wide variety of methods has been proposed to accomplish the tasks of network reconstruction and link prediction. Based on local or global measures, most of these have proved efficient mainly in the case of undirected and unweighted graphs. In some cases the proposed measures can be adapted to the directed and/or weighted cases but it is often challenging to find a method readily applicable to both scenarios and that can be used for network reconstruction and link imputation as well.
Here we propose a maximum entropy method for network reconstruction and link imputation, suitable for directed weighted networks. The comparison between our method and some existing ones shows the goodness of imputation, especially with respect to the AUC measure. Therefore our method results to be more versatile and always applicable, showing (slightly) better performances with respect to other standard methods.

2 Methods

The first step in order to both reconstruct a network and impute its missing links is to compute the probability of existence of a link between each couple of nodes. In the directed unweighted case, we solve a Shannon entropy constrained maximization problem using the Directed Binary Configuration Model (DBCM) [1] as a null model.The entropy to be maximized can be expressed in terms of the link probabilities as

$$S = -\sum_{i\neq j}[P_{i\to j}\log(P_{i\to j}) + (1-P_{i\to j})\log(1-P_{i\to j})]. \tag{1}$$

The probability of each link can be expressed in one of the two following ways

$$P_{i\to j} = \frac{x_i y_j}{1+x_i y_j} \quad \text{or} \quad P_{i\to j} = \frac{z s_i^{out} s_j^{in}}{1+z s_i^{out} s_j^{in}},\ \forall i = 1,...,N,\ \forall j \neq i \tag{2}$$

depending on the kind of accessible information (i.e. either binary or weighted). While in the first case the two vectors of fitnesses are obtained by solving the 2N coupled equations $\sum_{j\neq i}\frac{x_i y_j}{1+x_i y_j} = k_i^{out}$, $\sum_{j\neq i}\frac{x_j y_i}{1+x_j y_i} = k_i^{in}$ [1], in the second case the parameter z is obtained by imposing the constraint $L =< L >= \sum_i\sum_{j\neq i}P_{i\to j}$. [2, 3]

Link imputation Link prediction algorithms work by assigning a score to the subset of unobserved links and retaining only the ones characterized by the highest values of such score. The latter are defined in a variety of ways, each of which tries to capture the mechanism behind the network structure organization. In this paper, we employ the probability coefficients output by the DBCM as scores, in order to infer the L most probable, unobserved, connections. L is the number of links of the true underlying network (of adjacency A) and we assume to observe only a $\hat{L}$ of those. $\hat{A}$ denotes the adjacency matrix of the incomplete, observed, network. We proceed as follows:

- the probabilities, $P_{i\rightarrow j}$, are ordered in descending order;
- only the unobserved links are considered (the zeros of $\hat{A}$);
- we draw the $L-\hat{L}$ most probable links.

Network reconstruction Remarkably, our algorithm can be also used to define a deterministic reconstruction method, by applying the same procedure described above to the entire link set. In order to obtain a deterministic network reconstruction with the maximum entropy approach we need to assume that the total number of links, L, is known [10]. Then, the probabilities work as the score function of the reconstruction method.

3 Results

We tested the imputation method by randomly removing a given percentage of links from a known network and the use three indicators to evaluate the goodness of fit, and compare with the ones of other standard methods. The chosen indicators are: *area under the ROC curve* (AUC) [5], that represents the probability that a missing link is given a higher probability than a non-existent link; precision, the percentage of correctly recovered links over the total number of links; and accuracy, the ratio between the correct predictions over the total number of comparisons. Figure 1 shows the plot of the indicators for the maximum entropy approach (blue line), common neighbours (CN)[6] (red line), Jaccard index (J) [8] (green) and preferential attachment (PA) [7] (light blue), for the directed case, both binary (above) and weighted (below). It is to be noted that the extensions of CN, Jaccard and PA to the directed and weighted cases are not straightforward and make the score interpretation a more challenging task. That is because the use of strengths in place of degrees implies the loss of topological information. It is also relevant to notice that a better AUC score, provided by our method, is very important in the case in which the percentage of missing links might not be known exactly. In fact, unlike precision and accuracy, this measure does not depend on the number of links added but only on the probability assigned to the unobserved links.
In conclusion, we proposed a method based on Shannon entropy maximization for deterministic network reconstruction and link imputation. Remarkably, our method can be applied to any kind of network configuration, be it binary or weighted, contrarily to the majority of existing algorithms which have been tailored on binary, undirected networks and, as such, cannot be straightforwardly applied to different configurations. Another important feature of our method is that it allows also to estimate the weight of the imputed links. Although the quality of the estimation is still improvable, none of the compared method offers a weight estimation procedure.

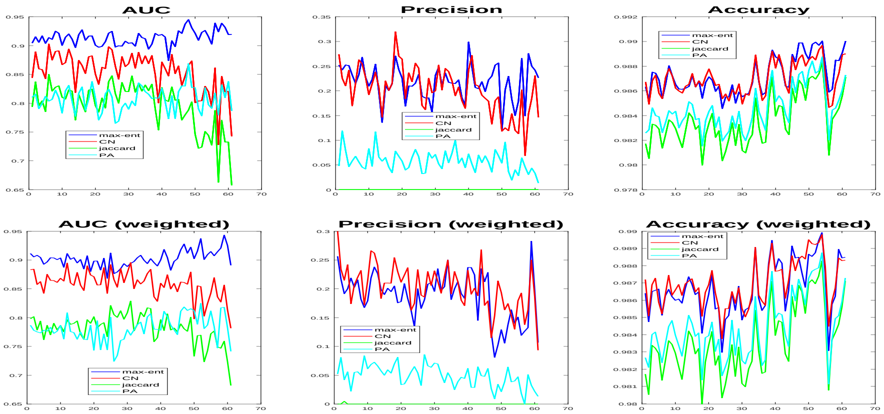


Fig. 1. Evaluation measures for the link imputation with 10% missing links. The network data are temporal snapshots from the Electronic Market for Interbank Deposits (E-mid) [9].

Future work

Among the different extensions to consider for future works, of interest are the ones that do not consider the observed data as ground truth. The simplest scenario is the one in which we consider that a fraction of links are to be removed as spurious. Even more interesting is link estimation performed in case of uncertain observation, as in [11]. This field of study is still under-explored and, as such, of great interest for future work.

References

1. Squartini, T., Fagiolo, G. Garlaschelli, D.: Randomizing world trade. I. A binary network analysis. Phys. Rev. E 84, 046117 (2011)
2. Cimini, G., Squartini, T., Garlaschelli, D., Gabrielli, A.: Systemic risk analysis on reconstructed economic and financial networks. Scientific reports 5 (2015).
3. Squartini, T., Cimini, G., Gabrielli, A., Garlaschelli, D.: Network reconstruction via density sampling. Applied Network Science (2017).
4. Mastrandrea, R., Squartini, T., Fagiolo, G., Garlaschelli, D.: Enhanced reconstruction of weighted networks from strengths and degrees. J. Soc. Ind. Appl. Math. 9(4), 533–543 (1961)
5. Hanely, J. A., McNeil, B. J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29.
6. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. Journal of the American society for information science and technology 58, 7 (2007)
7. Barabasi, A., Albert, R.: Emergence of scaling in random networks. Science 286 (1999).
8. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Socit Vaudoise des Sciences Naturelles 37 (1901)
9. De Masi, G., Iori, G., Caldarelli, G.:Fitness model for the Italian interbank money market. Phys. Rev. E 74 (2006).
10. Anand, K. et al. : The missing links: A global study on uncovering financial network structures from partial data. J. Financial Stability (2017).
11. Martin, T., Ball, B., Newman, M. E. J.: Structural inference for uncertain networks. Phys. Rev. E (2016).

Local-ring network automata and the impact of hyperbolic geometry in complex network link-prediction

Alessandro Muscoloni[1] and Carlo Vittorio Cannistraci[1,2,*]

[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[2] Brain bio-inspired computation (BBC) lab, IRCCS Centro Neurolesi "Bonino Pulejo", Messina, Italy

Methods for topological link-prediction are generally referred as global or local. The former exploits the entire network topology, the latter adopts only the immediate neighbourhood of the link to predict. Global methods are 'believed' to be the best performing. Is this common belief well-founded?

Stochastic-Block-Model (SBM) [1] is a global method believed as one of the best link-predictors and widely accepted as reference when new methods are proposed. But, our results suggest that SBM, whose computational time is high, cannot in general overcome the Cannistraci-Hebb (CH) network automaton model [2] that is a simple local-learning-rule of topological self-organization proved by multiple sources as the current best local-based and parameter-free deterministic rule for link-prediction [2]–[8]. In order to elucidate the reasons of this unexpected result, we formally introduce the notion of local-ring network automata models and their tight relation with the nature of common-neighbours' definition in complex network theory.

In addition, after extensive tests, we recommend Structural-Perturbation-Method (SPM) [9] as the new best global method baseline. However, even SPM overall does not outperform CH and in several evaluation frameworks we astonishingly found the opposite. In particular, CH was the best predictor for synthetic networks generated by the Popularity-Similarity-Optimization (PSO) model [10], and its performance in PSO networks with community structure [11] was even better than using the original internode-hyperbolic-distance as link-predictor. Interestingly, when tested on non-hyperbolic synthetic networks the performance of CH significantly dropped down indicating that this rule of network self-organization could be strongly associated to the rise of hyperbolic geometry in complex networks.

In conclusion, we warn the scientific community: the superiority of global methods in link-prediction seems a 'misleading belief' caused by a latent geometry bias of the few small networks used as benchmark in previous studies. Therefore, we urge the need to found a latent geometry theory of link-prediction in complex networks.

References

[1] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 52, pp. 22073–22078, 2009.

[2] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, "From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks," *Sci. Rep.*, vol. 3, no. 1613, pp. 1–13, 2013.

[3] Z. Liu, J. L. He, K. Kapoor, and J. Srivastava, "Correlations between Community Structure and Link Formation in Complex Networks," *PLoS One*, vol. 8, no. 9, 2013.

[4] L. Pan, T. Zhou, L. Lü, and C.-K. Hu, "Predicting missing links and identifying spurious links via likelihood analysis," *Sci. Rep.*, vol. 6, pp. 1–10, 2016.

[5] R. Pech, D. Hao, L. Pan, H. Cheng, and T. Zhou, "Link Prediction via Matrix Completion," *EPL*, no. 117, p. 38002, 2017.

[6] F. Tan, Y. Xia, and B. Zhu, "Link prediction in complex networks: A mutual information perspective," *PLoS One*, vol. 9, no. 9, 2014.

[7] T. Wang, H. Wang, and X. Wang, "CD-Based indices for link prediction in complex network," *PLoS One*, vol. 11, no. 1, pp. 5–7, 2016.

[8] W. Wang, F. Cai, P. Jiao, and L. Pan, "A perturbation-based framework for link prediction via non-negative matrix factorization," *Sci. Rep.*, vol. 6, no. December, p. 38938, 2016.

[9] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, "Toward link predictability of complex networks," *Proc. Natl. Acad. Sci.*, vol. 112, no. 8, pp. 2325–2330, 2015.

[10] F. Papadopoulos, M. Kitsak, M. A. Serrano, M. Boguna, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, pp. 537–540, 2012.

[11] A. Muscoloni and C. V. Cannistraci, "A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities," *arXiv:1707.07325 [physics.soc-ph]*, 2017.

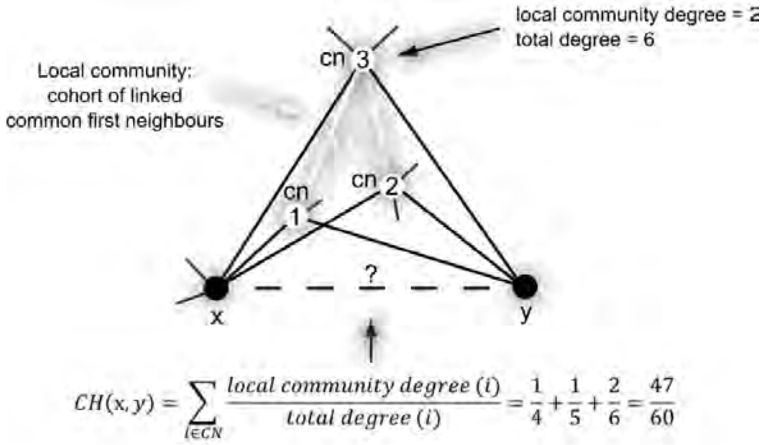

local community degree = 2
total degree = 6
Local community:
cohort of linked
common first neighbours
cn 3
cn 2
cn
1
?
x
y
$CH(x,y) = \sum_{i \in CN} \frac{local\ community\ degree\ (i)}{total\ degree\ (i)} = \frac{1}{4} + \frac{1}{5} + \frac{2}{6} = \frac{47}{60}$

Fig. 1. The figure shows an explanatory example for the topological link-prediction performed using the Cannistraci-Hebb (CH) rule. The two black nodes represent the seed nodes whose non-observed interaction should be scored with a likelihood. The three white nodes are the *common-neighbours* (CNs) of the seed nodes, further neighbours are not shown for simplicity. The cohort of linked common-neighbours forms the local-community and the links between them are called *local-community-links* (LCLs). Neighbours of the CNs that are neither the seed nodes nor in the local-community are indicated through chunks of outgoing links, which we named *external-local-community-links* (eLCLs). The mathematical formula for the CH index is reported, together with the detailed steps for computing the likelihood score for the link under analysis.

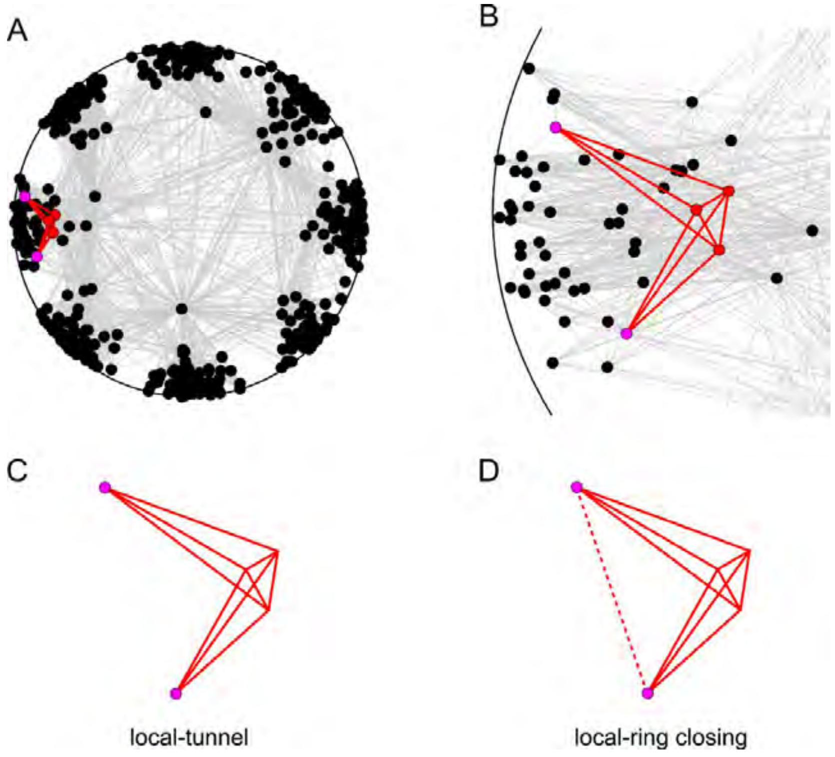


Fig. 2. The figure suggests a geometrical interpretation about how the CH network automaton model works in a monopartite topology. (A) Representation of a nPSO network [11] (γ = 3, m = 10, T = 0.1, N = 500, C = 8) in the hyperbolic space. In violet two non-adjacent nodes, whereas in red their CNs, the links from the two nodes to their CNs and the LCLs. (B) A zoom of the network in the region of two non-adjacent nodes considered for link formation. (C) Defining the *local-path* as the smallest possible path allowed between two non-adjacent nodes on a certain network topology (two-steps path in monopartite networks), the *local-tunnel* is the topological structure created by the ensemble of all the local-paths between the two non-adjacent nodes, plus the LCLs between CNs. The local-tunnel (which is formed in a hidden high-dimensional geometrical space that here, for simplifying the visualization, we project in the hyperbolic disk) provides a route of connectivity between the two non-adjacent nodes. (D) The addition of the link between the two non-adjacent nodes transforms the local-tunnel in a *local-ring* (*local-ring closing procedure*). The higher the number of CNs, the higher the volume of the local-tunnel. For each CN, the higher the number of LCLs in comparison to the eLCLs, the more the shape of the local-tunnel is well-defined and therefore its existence confirmed. Therefore, in link-prediction, CH estimates a likelihood that is proportional both to the volume of the local-tunnel and to the extent to which the local-tunnel exists.

Trajectory stability in the traveling salesman problem

Sergio Sánchez[1], Germinal Cocho[1], Jorge Flores[1], Carlos Gershenson[2], Gerardo Iñiguez[3], Carlos Pineda[4]

[1] Instituto de Física, Universidad Nacional Autónoma de México, 01000 CDMX, Mexico
[2] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, 01000 CDMX, Mexico
[3] Department of Computer Science, Aalto University School of Science, 00076 Aalto, Finland
[4] Faculty of Physics, University of Vienna, 1090 Wien, Austria
carlosp@fisica.unam.mx

Abstract. Two generalizations of the traveling salesman problem in which sites change their position in time are presented. The way the rank of different trajectory lengths changes in time is studied using the rank diversity. We analyze the statistical properties of rank distributions and rank dynamics and give evidence that the shortest and longest trajectories are more predictable and robust to change, that is, more stable [4].

1 Introduction

Many variations of the TSP have been analyzed in recent decades [2]. Previous research related to the TSP has focused mainly on producing algorithms to find shortest paths but, to our knowledge, the properties of longer trajectories have not been discussed. In the present work we study the statistical properties of all trajectories in two generalizations of the TSP with time-dependent sites: the TSP *with moving sites* (bTSP), where sites can be interpreted as 'boats' that move gradually in a region, and the TSP *with reallocation of sites* (rTSP), where sites move discontinuously across space. If we rank trajectories by their length, we can analyze how the properties of this ranking change in time with measures commonly used in the study of hierarchy dynamics in complex systems, such as the rank distribution $f(k)$ and the rank diversity $d(k)$ [1, 3].

2 Results

Consider the static TSP with N sites. We shall label each site by a, b, c, and so on. A *trajectory* is a closed path on these sites, so each trajectory can characterized by a non-unique string of site labels. There are $(N-1)!/2$ different trajectories, and the usual problem is to find the shortest one within this set. In this work we go further and rank each trajectory according to its length, so the shortest one has the highest rank ($k=1$), and the longest has the lowest rank ($k=k_{\max}=(N-1)!/2$).

We now study the problem of stability of trajectories in the TSP. Suppose the location of sites varies slowly over time, so that the salesman can assume a static scenario for each travel, but the shortest trajectory, and in fact the rank of all trajectories, might change if the configuration of sites is modified enough. A toy model that captures this situation is the bTSP. Assume all N sites are allowed to move within a 1×1 square as if they were boats. In the TSP with moving sites, the X and Y components of the velocity of each site are random with a uniform distribution between ± 1. The boats move with a constant velocity until they reach a confining wall, where they bounce elastically.

Let us consider a different time-dependent perturbation of the TSP, the rTSP. Here N sites are located in the unit square with a uniform random probability, and all trajectories are ranked as explained above. This is the “base” configuration. Then one site is chosen at random and reallocated to a random position in the unit square, after which trajectory ranks are calculated again. After several iterations, we can explore this time-dependent process with the rank diversity. Even though a continuous time dynamics does not exist as in the bTSP, we can still ask how many different trajectories occupy a given rank k, so the rank diversity is well defined.

Two useful generalizations of the bTSP and the rTSP are now considered, since it will allow us to relate the behavior of both models. Assume that only some sites move in the bTSP. That is, instead of all site moving in the plane, n move and $N-n$ are static. The cases analyzed before thus correspond to $n = N$. In a similar way, consider an rTSP where instead of reallocating a single site, n sites are moved. The case $n = N$ thus corresponds to a total reallocation of the system, while up to now we have only discussed the value $n = 1$. Diversity for these generalizations seems to behave in a similar way as for the case $n = 1$. In fact, for the bTSP one may even consider a single moving site and the previous conclusions still hold. We have obtained similar results for other variations of the bTSP, such as periodic boundary conditions, and different ways of choosing the initial conditions and velocities.

It is also useful to study the expected value of the stability area for different ranks, $\langle A \rangle_{\rho_k}$, as a function of the number of sites. There is a different scaling for the extremal and intermediate rankings, so we expect that the difference in areas seen for the case $N = 7$ is exponentially larger for bigger systems.

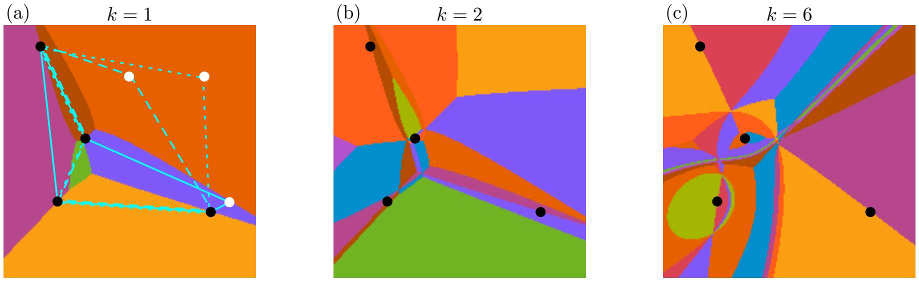


Fig. 1. Stability areas for different trajectories. (a) rTSP with $N = 5$, $n = 1$, and three possible reallocations of the white point. The shortest trajectory, $k = 1$, is indicated as a solid, dashed and dotted line, corresponding to the three reallocations. Regions of varying color correspond to different trajectories for $k = 1$ (see Figure **??**, top row). In panels (b) and (c) we show a plot with the same fixed sites as in (a), but for $k = 2$ and $k = 6$, respectively.

Consider the rTSP with $n = 1$. For a particular rank k, each position in the unit square yields a trajectory. We can thus 'paint' the unit square with colors corresponding to trajectories. We show examples with $N = 5$ for the shortest trajectory in Figure 1(a), and for $k = 2$ and 6 in Figure 1(b) and Figure 1(c), respectively. Diversity is the number of different colors in the picture divided by the number of observations. Notice that there is a qualitative difference between the cases $k = 1$ and $k = 6$. This indicates already that $d(1) < d(6)$ if a large ensemble is taken (so that errors due to finite sampling are small enough).

For the bTSP with $n = 1$, the moving boat will cover the whole unit square uniformly for most choices of the velocities. The condition for having a uniform covering is that the vertical and horizontal speeds are incommensurable, which always holds for this model. When this occurs, time averages yield the same value as space averages, i.e., the system is ergodic. For such long times the whole unit square will be visited, i.e. the moving site will visit all colored areas. Therefore, the bTSP has the same rank diversity as the rTSP when the sampling m is the same (otherwise they are related by a constant factor). For a larger number of moving sites, $n > 1$, a similar reasoning holds.

The bTSP and rTSP are comparable since both explore a fraction of the $2N$-dimensional configuration space. The bTSP explores a line of finite length in such a space, or a $2N$ dimensional hypercube embedded in the configuration space if the whole ensemble of velocities is considered. The rTSP, on the other hand, explores a $2n$ dimensional hyperplane embedded in the same configuration space. Overall, both models behave in a similar fashion.

Each realization of the TSP can be seen as a point in a $2N$-dimensional configuration space, where every pair of axis defines the coordinates of each particle. The optimization problem is then different for each point in the configuration space. We have analyzed the stability of the solutions of the TSP under changes of the location of the point defining the configuration. We have further shown that the stability properties are similar for the two time-dependent generalizations of the TSP considered here. We have also stated under what conditions the behavior of both models is identical. We thus expect that these results are applicable to other perturbations of the TSP that involve small variations in configuration space.

References

1. Cocho, G., Flores, J., Gershenson, C., Pineda, C., Sánchez, S.: Rank diversity of languages: Generic behavior in computational linguistics. PLoS ONE 10(4), e0121898 (04 2015),
2. Gutin, G., Punnen, A.: The Traveling Salesman Problem and Its Variations. Combinatorial Optimization, Springer US (2007), https://books.google.com.mx/books?id=pfRSPwAACAAJ
3. Morales, J.A., Sánchez, S., Flores, J., Pineda, C., Gershenson, C., Cocho, G., Zizumbo, J., Rodríguez, R.F., Iñiguez, G.: Generic temporal features of performance rankings in sports and games. EPJ Data Science 5(1), 33 (2016), http://dx.doi.org/10.1140/epjds/s13688-016-0096-y
4. Sánchez, S., Cocho, G., Flores, J., Gershenson, C., Iñiguez, G., Pineda, C.: Trajectory stability in the traveling salesman problem. ArXiv e-prints 1708.06945 (Aug 2017)

Rich-Club Ordering and the Dyadic Effect: Two Interrelated Phenomena

Matteo Cinelli[1], Giovanna Ferraro[1], and Antonio Iovanella[1]

Department of Enterprise Engineering, University of Rome "Tor Vergata",
Via del Politecnico, 1 - 00133 Rome, Italy.

1 Introduction

Rich-club ordering and the dyadic effect are two phenomena observed in complex networks that are based on the presence of certain patterns in the linkage of specific nodes. Rich-club ordering represents the tendency of highly connected and important elements to form tight communities with other central elements. The dyadic effect denotes the tendency of nodes that share a common property to be much more interconnected than expected. Herein, we consider the interrelation between these two phenomena, which until now have always been studied separately, providing a new formulation of the rich-club measures in terms of the dyadic effect. The reformulation allows us to improve the rich-club coefficient and to introduce certain measures, related to the analysis of the dyadic effect, which are useful in that they confirm the presence and relevance of rich-clubs in complex networks. Moreover, the introduced measures provide a baseline for the evaluation of the rich-club size, that is a open issue in the study of complex networks, in a computationally efficient way.

2 Involved Measures

The rich-club coefficient can be written as: $\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)}$. In the equation, $E_{>k}$ is the number of edges among the $N_{>k}$ nodes having degree d_i higher than a given value k and $\frac{N_{>k}(N_{>k}-1)}{2}$ is the maximum possible number of edges among the $N_{>k}$ nodes. Therefore $\phi(k)$ measures the fraction of edges connecting the $N_{>k}$ nodes out of the maximum number of edges they might possibly share.

When we consider the dyadic effect, we have to take into account the network nodes together with their metadata (or characteristics) in a binary form thus, we refer to a given node characteristic c_i, which can assume the values 0 or 1 for each $i \in N$. Consequently, N can be divided into two subsets: the set of n_1 nodes with characteristic $c_i = 1$, the set of n_0 nodes with characteristic $c_i = 0$; thus, $N = n_1 + n_0$. We distinguish then three kinds of *dyads*, i.e. edges and their two end nodes, in the network: $(1-1)$, $(1-0)$, and $(0-0)$. We label the number of each dyad in the graph as m_{11}, m_{10}, m_{00}, respectively. However, the number of each kind of dyad cannot assume arbitrary values and in the case of m_{11} we exploit an upper bound introduced in [1] that will be useful in improving the coefficient $\phi(k)$ using the dyadic effect notation. The upper bound is written as:

$$UBm_{11} = min\left(M, \binom{n_1}{2}, \left\lceil \sum_{i \in D_G^H(n_1)} \frac{\min(d_i, n_1 - 1)}{2} \right\rceil \right) \quad (1)$$

It is based on the fact that large cliques are rare substructures in sparse networks and therefore it exploits the degree sequence D_G in order to check whether the network G can actually contain a complete subgraph of size n_1. If not, the densest hypothetical substructure that could be realized using the first elements (those with the highest degree) of the degree sequence of G is taken into account.

3 Reformulation and Results

Using the quantities from above, we can express the rich-club coefficient in terms of the dyadic effect. If we write: $c_i = 1$ if $d_i > k$ and $c_i = 0$ if $d_i \leq k$ we immediately obtain $N_{>k} = n_1$ and $N_{\leq k} = n_0$. Thus, the nodes with degree higher (lower/equal) than a certain threshold k correspond to those having the characteristic $c_i = 1$ ($c_i = 0$). Since $E_{>k}$ is defined as the number of edges between nodes $N_{>k} = n_1$, we can consequently write $E_{>k} = m_{11}$, $E_{\leq k} = m_{00}$ and consequently $\overline{E}_{>k} = M - E_{>k} - E_{\leq k} = M - m_{11} - m_{00} = m_{10}$.

Therefore, we can reformulate $\phi(k)$ as:

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)} = \frac{m_{11}}{\binom{n_1}{2}} \tag{2}$$

Since the upper bound UBm_{11} includes the denominator of $\phi(k)$ in its formula, it is therefore possible to introduce the improved version of $\phi(k)$ written as:

$$\phi(k)^{new} = \frac{m_{11}}{UBm_{11}} \tag{3}$$

By using combinatorial arguments the upper bound UBm_{11} captures the densest realizable substructure when a clique of n_1 nodes is not realizable within the given D_G; in summary, $\phi(k) \leq \phi(k)^{new}$ when $\binom{n_1}{2} \geq UBm_{11}$. Thus, $\phi(k)^{new}$ identifies the presence of the rich-club normalizing the actual amount of links among the rich nodes not over a clique of size n_1, but on a sparser subgraph of the same size deriving from the degree sequence of the considered network. This imply that $\phi(k)^{new}$ is a better measure with respect to $\phi(k)$ as it results to be more tailored for the network under investigation. Moreover, the detection of rich-club ordering (i.e. the computation of the coefficient $\phi(k)_{norm}$ [3]) can be performed by using equally $\phi(k)$ or $\phi(k)^{new}$, as shown in [2]. As an example, we compute the two coefficients $\phi(k)$ and $\phi^{new}(k)$ on Internet at AS level on which the presence of a rich-club has been observed for the first time [6].

By looking at Figure 1, we note that, up to a certain point, our reformulation outperforms the older one because of the impossibility to realize a clique of size n_1 using the degree sequence of the considered network. Conversely, the two coefficients become equal when a clique is realizable (i.e. for very low values of n_1 as shown later) and thus the network structure allows the presence of a dense rich-club. It is worth mentioning that the rich-club phenomenon is studied in the case of sparse networks and that the situations in which a dense rich-club is likely to be present are those of interest in the study of real networks as the rich-club structural properties are able to provide insights about local and global aspects of the whole system [6].

In more detail, the curves $\phi(k)$ and $\phi(k)^{new}$ encounter each other in correspondence to the degree value k, which is necessary to overcome in order to guarantee a certain

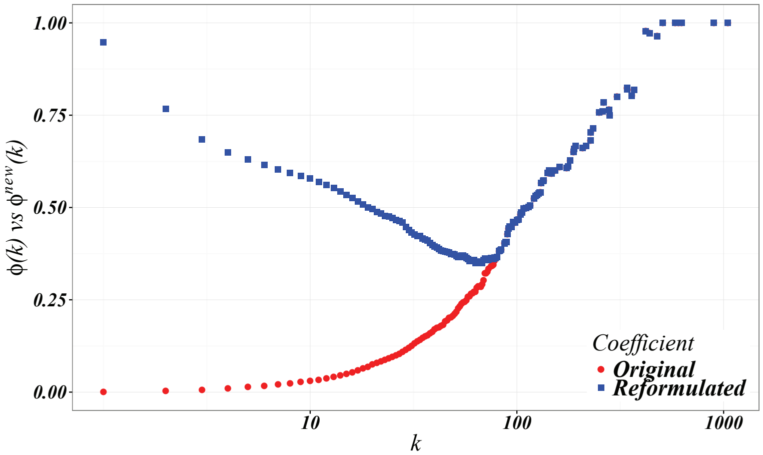


Fig. 1. Curves related to $\phi(k)$ vs $\phi(k)^{new}$ and $\delta(k)$ for the AS-level Internet network.

likelihood of finding a complete subgraph of size n_1. As such, the junction point of the curves related to the two coefficients can thus be interpreted as a first indicator of the rich-club size. In the Internet network ($N = 11461$, $M = 32730$), the two curves meet when $k = 83$ and in such a network the number of nodes with degree higher than k is $N_{>k} = n_1 = 84$. Therefore, if we consider the junction point as a rich-club size indicator, we would obtain a subgraph with $n_1 = 84$ nodes, i.e. a subgraph whose size is 0.7% ($\frac{n_1}{N} = 0.0073$) with respect to the whole network. Thus our measures are able to estimate the size of the rich-club around the 1% of the network nodes which is a value in accordance with the empirical evidence [5] [6].

As previously mentioned, the issue of knowing whether a certain network possesses a rich-club is important for many reasons, including factors relating to either its functional or topological role as shown, for instance, by the extensive utilization of this measure in biology and especially in neuroscience [4].

References

1. Cinelli, M., Ferraro, G., Iovanella, A.: Structural bounds on the dyadic effect. Journal of Complex Networks 5(5), 694–711 (2017), + http://dx.doi.org/10.1093/comnet/cnx002
2. Cinelli, M., Ferraro, G., Iovanella, A.: Rich-club ordering and the dyadic effect: Two interrelated phenomena. Physica A: Statistical Mechanics and its Applications 490(Supplement C), 808 – 818 (2018), http://www.sciencedirect.com/science/article/pii/S037843711730852X
3. Colizza, V., Flammini, A., Serrano, M.A., Vespignani, A.: Detecting rich-club ordering in complex networks. Nature physics 2(2), 110–115 (2006)
4. Van Den Heuvel, M.P., Sporns, O.: Rich-club organization of the human connectome. Journal of Neuroscience 31(44), 15775–15786 (2011)
5. Xu, X.K., Zhang, J., Small, M.: Rich-club connectivity dominates assortativity and transitivity of complex networks. Physical Review E 82(4), 046117 (2010)
6. Zhou, S., Mondragón, R.J.: The rich-club phenomenon in the internet topology. IEEE Communications Letters 8(3), 180–182 (2004)

Core Decomposition of Uncertain Graphs Using Representative Instances

Damien Seux[1], Fragkiskos D. Malliaros[2], Apostolos N. Papadopoulos[3], and
Michalis Vazirgiannis[1]

[1] Computer Science Laboratory, École Polytechnique, France
{damien.seux, michalis.vazirgiannis}@polytechnique.edu
[2] Center for Visual Computing, CentraleSupélec and Inria, France and UC San Diego, USA
fmalliaros@eng.ucsd.edu
[3] Department of Informatics, Aristotle University of Thessaloniki, Greece
apostol@csd.auth.gr

1 Introduction and Problem Statement

In many real-world applications, the corresponding graphs are inherently associated with *uncertainty*, which can be due to various reasons, such as uncertainty introduced by the data-collection process or for privacy-preserving reasons. For example, in the case of protein-protein interaction networks (PPI) in the domain of biology, each node corresponds to a specific protein and the edges capture information about the interaction of two proteins. Since in many cases those interactions are indicated either by noisy laboratory experiments or by prediction algorithms based on features of the proteins (instead of being actually observed), a level of uncertainty is introduced in the edges of the graph. This uncertainty can be captured by the model of *uncertain* or *probabilistic* graphs, where each edge is associated with a probability of existence.

In this work, we are interested in a widely applied graph analytics tool, namely the one of *k-core decomposition* [4]. Let H be a subgraph of graph G. Subgraph H is defined to be a *k-core* of G, denoted by G_k, if it is a maximal connected subgraph of G in which all vertices have degree at least k. Based on that, vertex i has *core number* $\text{core}_G(i) = k$, if it belongs to a k-core but not to any $(k+1)$-core. Due to its simplicity and computational efficiency, the k-core decomposition has been applied in many domains, including community detection and identification of influential spreaders in social networks. Then, the following questions arise, which also describe the goals of this work: *how to define the concept of core decomposition in uncertain graphs and how to efficiently compute it?* Bonchi et al. [2] proposed an extension of the k-core decomposition to uncertain graphs which requires that the probability that each vertex v within the core subgraph H has degree at least k, is greater than or equal to a parameter η. Nevertheless, this definition has two main weaknesses: (i) an extra probability threshold η is required in order to define the core structure – making the resulting decomposition dependent on this user-defined parameter; (ii) the increased computational cost for performing the decomposition. Based on that, our goal is to define a simple-yet-effective core decomposition of uncertain graphs. To do so, we consider the *expected degree* of each vertex in the uncertain graph, and in particular the concept of *representative instance*.

2 Cores in Probabilistic Graphs

Let $\mathscr{G} = (V, E, p)$ be an uncertain graph, where $p : E \rightarrow (0, 1]$ is a function that assigns probabilities to the edges of the graph. A widely used approach to analyze uncertain graphs is the one of *possible worlds*, where each possible world constitutes a deterministic realization of $\mathscr{G}$. According to this model, an uncertain graph $\mathscr{G}$ is interpreted as a set $\{G = (V, E_G)\}_{E_G \subseteq E}$ of $2^{|E|}$ possible deterministic graphs [3]. Let $G \sqsubseteq \mathscr{G}$ indicates that G is a possible world of $\mathscr{G}$. Then, the probability that $G = (V, E_G)$ is observed as a possible world of $\mathscr{G}$ is given by $\Pr(G|\mathscr{G}) = \prod_{e \in E_G} p(e) \ \prod_{e \in E \setminus E_G} (1 - p(e))$. In our approach, we are using this general framework to derive an analogous of the k-core decomposition in uncertain graphs. In particular, we use the property of *expected degree* $[d](v)$ of each node $v \in \mathscr{G}$, leading to the concept of uncertain $[k]$-core decomposition.

Definition 1 (Uncertain $[k]$-core). *Given an uncertain graph $\mathscr{G} = (V, E, p)$, the uncertain $[k]$-core of $\mathscr{G}$ is the maximal subgraph $\mathscr{H} = (C, E|C, p)$ such that each vertex $v \in C$ has expected degree at least k in $\mathscr{H}$, where $k \in \mathbb{R}^+$.*

According to this definition, in order to compute the decomposition, we can extract a *deterministic representative instance* $G \sqsubseteq \mathscr{G}$ that preserves the expected degree, i.e., the degree of any vertex $v \in G$ to be as close as possible to the expected degree of $v \in \mathscr{G}$ – therefore, casting the problem to a weighted version of the k-core decomposition on deterministic graphs. That way, the proposed algorithm comprises of two phases: (i) extraction of a representative instance of the uncertain graph; (ii) apply a modified version of the k-core decomposition, suitable for fractional degree values.

For the first step of the algorithm that converts the uncertain graph to a deterministic one by preserving the expected degree, we rely on conversion algorithms that aim at minimizing the *discrepancy* of each vertex of the graph [3]. In particular, the discrepancy $\text{dis}_G(v)$ of a vertex v in the representative instance $G \sqsubseteq \mathscr{G}$, is defined as the difference between the degree in the representative instance and expected degree in the uncertain graph, i.e., $\text{dis}_G(v) = d(v) - [d](v)$. As the existence of the edges of the graph are independent of each other, the expected degree $[d](v)$ of a vertex $v \in V$ is the sum of the probabilities of the incident edges, i.e., $[d](v) = \sum_{e=(v,u) \in E} p(e)$. The overall discrepancy of the representative instance $G \sqsubseteq \mathscr{G}$ is defined as $\Delta(G) = \sum_{v \in V} |\text{dis}_G(v)|$. Then, the problem of finding a "good" representative instance G^* can be expressed as a minimization optimization problem: $G^* = \arg\min_{G \sqsubseteq \mathscr{G}} \Delta(G)$.

After extracting an *average degree-preserving* representative instance of the uncertain graph, the core number of a vertex is not an integer anymore but a real number. Thus, the second phase of the proposed technique consists of a modified k-core decomposition algorithm that operates on a *deterministic graph* with fractional node degrees.

3 Experimental Results and Discussion

We have performed preliminary experiments on a co-authorship network (DBLP) derived from DBLP (http://dblp.uni-trier.de), that consists of $404,892$ nodes and $1,422,263$ edges. Since the DBLP dataset is not inherently uncertain, we are using a method to convert the graph to uncertain by examining the similarity between the neighborhood

Table 1. Properties of the $[k]$-core decomposition on the `DBLP` graph.

Correlation	*Partial*	*Full*
τ_B	0.66	0.12
r	0.79	−0.02

of two nodes – towards computing the probability of being linked. In particular, we have applied the *Jaccard similarity coefficient* to quantify the similarity of two nodes, following two different approaches. The first one, denoted by *Partial*, is adding a probability of existence only to current edges of the graph. However, it does not allow to consider more pairs of nodes, than the already existing edges in the graph. The second approach, denoted by *Full*, is computing the Jaccard similarity coefficient for all pairs of nodes. To overcome the complexity of examining all possible pairs, we restrict our interest to pairs that have at least one neighbor in common. Then, we compute the core decomposition on both the original graph and the uncertain one, and examine the correlation among them using the Kendall rank correlation coefficient τ_B (which measures how much the ranking output is the same in both cases), and Pearson's correlation r (which measures how much the core numbers are linearly related).

Table 1 depicts the results. As we observe, computing the similarity only for existing edges retains the structural properties of the decomposition in both cases. Nevertheless, notice that in this case the original graph can easily be recovered from the uncertain one (for example, when obfuscating the graph for privacy preserving reasons making it uncertain, this approach is not possible). However, computing the probabilities for all pair of nodes (*Full*), erases all the structural information of cores (both correlation measures are close to zero).

Currently, we are working towards examining practical applications of the proposed $[k]$-core decomposition in uncertain graphs. For example, in the `DBLP` graph, it would be interesting to conduct an exploratory study comparing the authors (i.e., nodes) belonging to the maximal core subgraph extracted by the algorithms on the deterministic and uncertain graphs respectively. Moreover, we plan to examine the performance of the high core number nodes detected by the proposed decomposition, in the task of influence maximization.

References

1. Batagelj V., Zaversnik M.: An $\mathcal{O}(m)$ algorithm for cores decomposition of networks. arXiv (2003)
2. Bonchi, F., Gullo, F., Kaltenbrunner, A., Volkovich, Y.: Core decomposition of uncertain graphs. The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD), pp. 1316–1325 (2014)
3. Parchas, P., Gullo, F., Papadias, D., Bonchi, F.: Uncertain graph processing through representative instances. ACM Trans. Database Syst., 40(3):20:1–20:39 (2015)
4. Seidman, S. B.: Network Structure and Minimum Degree. Social Networks, 5:269–287 (1983)

Part IX

Network Analysis and Measures

Kemeny's constant and the effective graph resistance

Xiangrong Wang[1], Johan L. A. Dubbeldam[1], and Piet Van Mieghem[1]

Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, P.O Box 5031, 2600 GA Delft, The Netherlands
x.wang-2@tudelft.nl

1 Introduction

The effective graph resistance R_G, also called Kirchhoff index, characterizes the resistance distance [4] between nodes in an electrical network and can be computed by $R_G = N\sum_{i=1}^{N-1}\frac{1}{\mu_i}$, where μ_i is the i-th eigenvalue of the Laplacian matrix Q of a graph G. Studies [2,7,6] relate the effective graph resistance and the trace of the pseudo-inverse Laplacian $Q^\dagger$ as $R_G = N\text{trace}(Q^\dagger) = N\sum_{j=1}^{N}\left(Q^\dagger\right)_{jj}$.

In complex networks, represented by graphs, the effective graph resistance characterizes the difficulty of transport in a network. As a robustness indicator, the effective graph resistance allows to compare graphs and is applied in improving the robustness of complex networks, especially against cascading failures in electrical networks [1,8].

Let P denote the transition probability matrix of a finite, irreducible Markov Chain and the steady state probability vector π and the all-one vector u satisfying $Pu = u$ and $\pi^T P = \pi^T$. It is shown [3] that the inverse $Z \equiv \left(I - P + gh^T\right)^{-1}$ exists if any two column vectors h and g have nonzero scalar products $h^T u$ and $\pi^T g$. The Kemeny constant is defined, in terms of the trace of the matrix Z, as

$$K(P) \equiv \text{trace}\,(Z) - \pi^T Zu$$

For a given transition probability matrix P and with $h^T g = 1$, the Kemeny constant $K(P)$ is the same regardless of the choice of the matrix Z.

Kemeny's constant and its relation to the effective graph resistance has been established for regular graphs by Palacios et al. [5]. Based on the Moore-Penrose pseudo-inverse of the Laplacian matrix, we derive a new closed-form formula and deduce upper and lower bounds for the Kemeny constant. Furthermore, we generalize the relation between the Kemeny constant and the effective graph resistance for a general connected, undirected graph.

2 Results

The adjacency matrix A of a graph G with N nodes and L links is an $N \times N$ symmetric matrix with elements a_{ij} that are either 1 or 0 depending on whether there is a link between nodes i and j or not. The Laplacian matrix Q of G is an $N \times N$ symmetric matrix $Q = \Delta - A$, where $\Delta = \text{diag}(d_i)$ is the $N \times N$ diagonal degree matrix with the elements $d_i = \sum_{j=1}^{N} a_{ij}$. Let $d = (d_1,\ d_2,\ \ldots,\ d_N)$ denote the degree vector for a graph

G. Assume the transition probability matrix $P = \Delta^{-1}A$, we derive two closed-form formulas for the Kemeny constant. One expression is

$$K(\Delta^{-1}A) = \zeta^T d - \frac{d^T Q^\dagger d}{2L} \tag{1}$$

where the column vector $\zeta = \left(Q^\dagger_{11},\ Q^\dagger_{22},\ \ldots\ Q^\dagger_{NN}\right)$ and another expression, in terms of the effective resistance matrix Ω, is

$$K(\Delta^{-1}A) = \frac{d^T \Omega d}{4L} \tag{2}$$

where $\Omega = (\omega_{ij})$ and each element ω_{ij} represents the effective resistance between nodes i and j.

The Kemeny constant $K(\Delta^{-1}A)$ in (1) is upper and lower bounded by

$$\zeta^T d - \frac{\mathrm{Var}\,[D]}{E\,[D]\,\mu_{N-1}} \leq K(\Delta^{-1}A) \leq \zeta^T d - \frac{\mathrm{Var}\,[D]}{E\,[D]\,\mu_1} \tag{3}$$

where D is the random variable of the degree in a graph, and $E\,[D]$, $\mathrm{Var}\,[D]$ are the average and the variance of the degree.

We numerically evaluate the upper and lower bounds in (3) for various random graphs. In Figure 2, we present the accuracy of the bounds for (a) Erdős-Rényi graphs (ER) with $N = 500$ nodes, link density $p = 2p_c$, where $p_c = \frac{\log(N)}{N}$ is the connectivity threshold; (b) Barabási-Albert graphs (BA) with $N = 500$ and the average degree $d_{av} = 6$. For each class of random graphs, we generate 10^5 graph instances and the probability density functions for the Kemeny constant $K(\Delta^{-1}A)$ and the bounds are plotted. The upper bound deviates on average 0.01% and 0.04% of the numerical value of the Kemeny constant in ER random graphs and BA graphs, respectively. The lower bound is slightly less accurate compared to the upper bound, with 0.05% and 0.8% of difference in ER and BA graphs. Hence, the simulation results show that the upper and lower bounds in (3) are a good approximation for $K(\Delta^{-1}A)$.

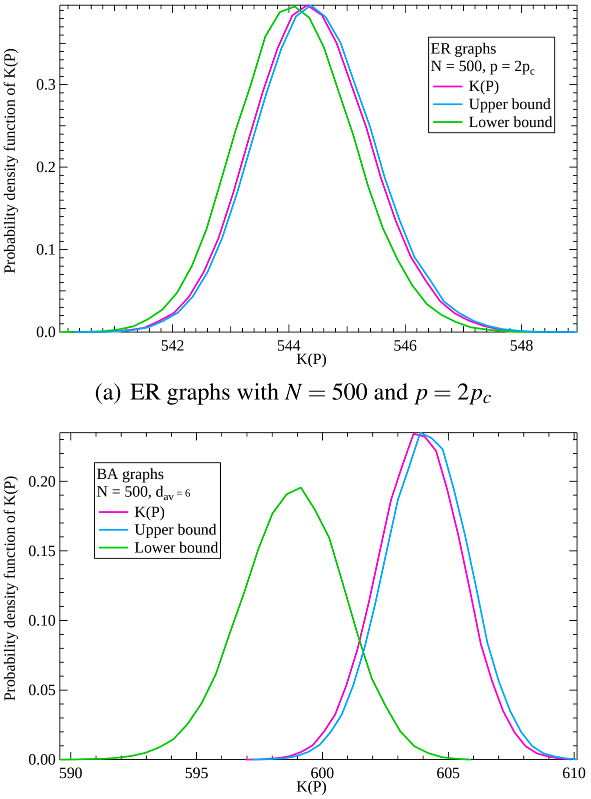
The relation between the Kemeny constant $K(\Delta^{-1}A)$ and the effective graph resistance R_G is generalized to an undirected, connected graph as

$$\frac{d_{min}R_G}{N} - \frac{d^T Q^\dagger d}{2L} \leq K(\Delta^{-1}A) \leq \frac{d_{max}R_G}{N} \tag{4}$$

where $d_{\min}$ and $d_{\max}$ is the minimum and the maximum degree in graph G, respectively.

Summary. In this extended abstract, we generalize the relation between the Kemeny constant and the effective graph resistance, which was known for regular graphs, to general connected, undirected graphs. By deriving a new closed-form formula (1), we provide a new approach to compute the Kemeny constant via the pseudo-inverse of the Laplacian matrix. Moreover, we show that for general graphs the Kemeny constant can be tightly upper and lower bounded by (3).

We appreciate Eric J.E.M. Pauwels and Joost Berkhout for inspiring discussions. This work is part of the research program complexity in logistics with project number 439.16.107, which is financed by the Netherlands Organization for Scientific Research (NWO).



(a) ER graphs with $N = 500$ and $p = 2p_c$

(b) BA graphs with $N = 500$ and $d_{av} = 6$

Fig. 1. Accuracy of the upper and lower bounds for the Kemeny constant.

References

1. Ellens, W., Spieksma, F.M., Van Mieghem, P., Jamakovic, A., Kooij, R.E.: Effective graph resistance. Linear algebra and its applications 435(10), 2491–2506 (2011)
2. Gutman, I., Xiao, W.: Generalized inverse of the Laplacian matrix and some applications. Bulletin: Classe des sciences mathématiques et natturalles 129(29), 15–23 (2004)
3. Kemeny, J.G.: Generalization of a fundamental matrix. Linear Algebra and its Applications 38, 193–206 (1981)
4. Klein, D.J., Randić, M.: Resistance distance. Journal of mathematical chemistry 12(1), 81–95 (1993)
5. Palacios, J.L.: On the Kirchhoff index of regular graphs. International Journal of Quantum Chemistry 110(7), 1307–1309 (2010)
6. Ranjan, G., Zhang, Z.L.: Geometry of complex networks and topological centrality. Physica A: Statistical Mechanics and its Applications 392(17), 3833 – 3845 (2013)
7. Van Mieghem, P., Devriendt, K., Cetinay, H.: Pseudoinverse of the Laplacian and best spreader node in a network. Physical Review E 96(3), 032311 (2017)
8. Wang, X., Pournaras, E., Kooij, R.E., Van Mieghem, P.: Improving robustness of complex networks via the effective graph resistance. The European Physical Journal B 87(9), 221 (2014)

Open Networks from Within: From Categorical Network Theory to New Centrality Measures of Nodes as Input or Output

Taichi Haruna[1]

Department of Information and Sciences, Tokyo Woman's Christian University, Tokyo, Japan
tharuna@lab.twcu.ac.jp

Categorical network theory provides a general framework to study open networks, namely, networks with explicit input and output nodes such as electrical circuits, signal flow diagrams, chemical reaction networks and so on [1]. Its primary focus is on the behavior of open networks that are determined by the relation between inputs and outputs. Thus the internal structure of networks is black-boxed in categorical network theory. On the other hand, it goes without saying that analysis of network structure is one of the most important issues in network science [3]. In this study, we try to bridge these two approaches to networks. In particular, we show that new centrality measures of nodes as input or output in directed networks can be obtained by internalizing the idea of open networks: each node within a network has input and output parts. This is a natural supposition when we look at real-world networks. By formalizing this idea in category theory [7], we can show that the notion of lateral path between two *arcs* (Fig. 1) is identified as the category-theoretic universal structure with respect to the above idea of "internal openness" of each node in directed networks. This result is a recapitulation of the result obtained in [4], but with the new interpretation.

Lateral paths between two *nodes* have been used to identify the bipartite community structure in directed networks [2]. Nodes included in the source or target sides of an identified bipartite community could be interpreted as the input or output parts of a given directed network, respectively. In contrast, here we use lateral paths between two *arcs* to define a centrality of nodes as input or output in a directed network. The idea is that if a node is the source or target of arcs in many lateral paths, it would be potentially important as input or output, respectively. One way to quantify importance of nodes as input or output in this sense is to calculate the betweenness centrality of each arc with respect to lateral paths and project it to its source node or its target node, respectively: Let A be the set of arcs of a given directed network. The lateral betweenness centrality of an arc f is [4]

$$\mathrm{LBC}_f = C \sum_{g,h\in A,\ l_{gh}>0} \frac{l_{gh}^f}{l_{gh}}, \tag{1}$$

where l_{gh} is the number of shortest lateral paths between g and h, l_{gh}^f is the number of shortest lateral paths between g and h that pass through f and $C = \sum_{g,h\in A, l_{gh}>0}(d_{gh}+1)$ is the normalization constant such that $\sum_{f\in A}\mathrm{LBC}_f = 1$ where d_{gh} is the length of the shortest lateral paths between g and h. The input or output betweenness centrality (IBC

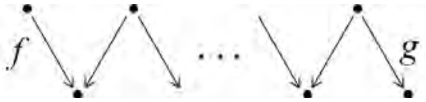


Fig. 1. Lateral paths are defined between two arcs, not between two nodes. A lateral path between two arcs f and g is a sequence of arcs between them such that successive arcs have a common target node or source node alternately.

or OBC) of a node x are defined as

$$\mathrm{IBC}_x = \sum_{s(f)=x} \mathrm{LBC}_f \tag{2}$$

or

$$\mathrm{OBC}_x = \sum_{t(f)=x} \mathrm{LBC}_f, \tag{3}$$

respectively. Here, $s(f)$ or $t(f)$ indicate the source or target node of f, respectively.

In Fig. 2 we show IBC and OBC of nodes in three biological networks. IBC and OBC of each node are plotted against its out-degree and in-degree, respectively. They are positively correlated with the corresponding degree. However, in the original biological networks, we can identify several nodes with IBC or OBC values that are significantly larger than those of randomized networks with degree-preservation. For example, in the neuronal network of *C. elegans*, the nodes with significantly large IBC values are neurons mediating avoidance behavior from chemicals or neurons involved in thermosensation (Fig. 2 (a)). In the food web of Florida Bay, five out of seven phytoplankton taxa are identified as having significantly large IBC values although they have relatively low out-degrees (Red squares around IBC ≈ 0.01 in Fig. 2 (e)). These initial tests suggest that IBC and OBC are potentially useful to capture the importance of nodes as input or output that cannot be assessed by simply calculating the out-degree or in-degree of nodes, respectively.

References

1. Baez, J. C., Coya, B., Rebro, F.: Props in network theory. arXiv:1707.08321.
2. Crofts, J. J., Estrada, E., Higham, D. H., Taylor, A.: Mapping directed networks. Electron. Trans. Numer. Anal. 37, 337 – 350 (2010)
3. Estrada, E.: The Structure of Complex Networks: Theory and Applications. Oxford University Press (2012)
4. Haruna, T.: Theory of interface: Category theory, directed networks and evolution of biological networks. BioSystems 114, 125 – 148 (2013)
5. Kaiser, M., Hilgetag, C. C.:Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. PLoS Computat. Biol. 2, e95 (2006)
6. Shen-Orr, S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat. Genet. 31, 64 – 68 (2002)
7. Spivak, D. I.: Category Theory for the Sciences. The MIT Press (2014)
8. Ulanowicz, R., Bondavalli, C., Egnotovich., M.: Network analysis of trophic dynamics in south Florida ecosystem, fy 97: The Florida bay ecosystem. Ref. No.[UMCES] CBL 98-123. Chesapeake Biological Laboratory, Solomons, MD 20688-0038, USA (1998)

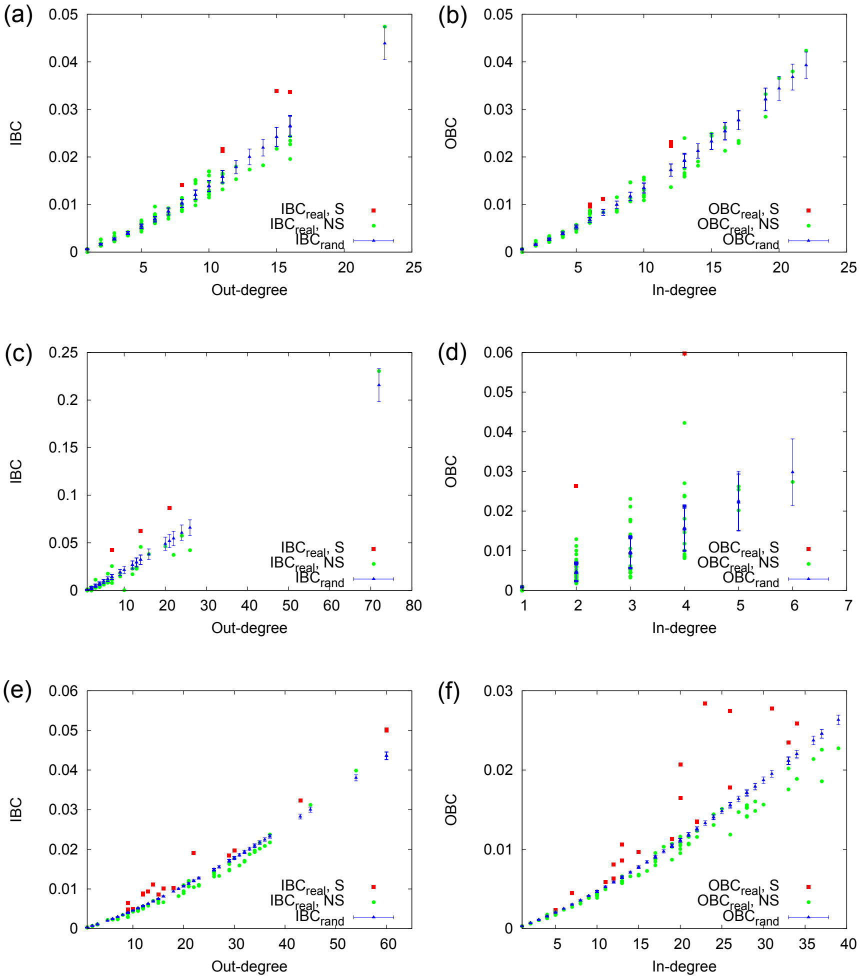


Fig. 2. Input betweenness centrality (IBC) and Output betweenness centrality (OBC) of each node in three biological networks are plotted against its out-degree and in-degree, respectively. The blue triangles are the average value over 1000 randomized networks with degree-preservation. The error bars represent the standard deviation. The red squares represent nodes whose IBC or OBC are judged to be significantly larger than those of the corresponding nodes in degree-preserved random networks by the following procedure, respectively. The green points are the other non-significant nodes. Let b denote IBC or OBC of a given node. The p-value of b is calculated from its z-score if the distribution of b in the null model can be approximated by a normal distribution. Practically, we tested this by the Kolmogorov-Smirnov test. If the p-value of the KS test is greater than 0.10, we adopted the normal approximation. Otherwise, it is simply the proportion of the degree-preserved randomization trials in which b exceeds the original value. Further, we applied the Benjamini-Hochberg-Yekutieli procedure at level 0.05 for the multiple comparisons correction. (a), (b) The neuronal network of *C. elegans* consisting of 131 neurons in the rostral ganglia [5]. (c), (d) The gene regulatory network of *E. coli* consisting of 328 genes (operons) [6]. (e), (f) The food web of Florida Bay consisting of 121 taxa [8].

Using network science to study the orthographic language network of English

Cynthia S. Q. Siew[1]

[1]Department of Psychology, University of Warwick, Coventry, United Kingdom
cynsiewsq@gmail.com

1 Introduction

Within the cognitive sciences, network science has been applied to study the structure of the mental lexicon, the part of long-term memory where all the words a person knows is stored[1]. The mental lexicon can be viewed as a language network, where nodes represent words and edges represent relationships between words. Words can be related to other words in different ways—semantically (i.e., a word's meaning; cat-dog), phonologically (i.e., the sounds of words; /k@t/-/h@t/), and orthographically (i.e., a word's spelling; 'cat'-'cap'). Past work has shown that semantic[2] and phonological[3] language networks have a small-world structure and that the structure of these networks influences various aspects of language processing—such as language acquisition[4] and spoken word recognition[5].

However, to date, not much is known about the *orthographic* language network, where edges in the network represent orthographic similarity relationships between words. Conceptualizing lexical representation as an orthographic network will build on previous psycholinguistic work demonstrating that orthographic similarity among words affects reading speeds[6] by providing new ways of quantifying and investigating the orthographic similarity structure of language. The present project aims to (1) construct an orthographic language network and analyze its overall network structure, and (2) determine if the structure of the orthographic language network influences visual word recognition performance.

2 Method

40,468 English words were obtained from the English Lexicon Project[7] (http://elexicon.wustl.edu/), a database of lexical and behavioral data. A link was placed between any two words that differed by a Levenshtein edit distance of 1 (i.e., whether the first word could be transformed into the second via the substitution, addition, or deletion of one letter[8]). For instance, 'cat' would be connected to 'hat', 'chat', and 'at'. The resulting orthographic language network consisted of 40,468 nodes and 41,514 edges. The sparseness of the network was due to the large proportion of nodes that either did not connect to any other nodes (40.74%) or found in smaller connected components (sizes range from 2 to 34; 31.17%). The largest connected component (LCC) of the orthographic language network consisted of 11,365 nodes and 32,759 edges. The LCC had an average degree <*k*> of 5.766, average clustering coefficient of 0.273, average path length of 8.78, and network diameter *D* of 31. The degree distribution was best approximated by a power-law distribution with an exponent of 1.77. Overall, the LCC of the orthographic network is scale-free and has a small-world structure, as characterized by a small average path length and large average clustering coefficient relative to a comparably sized random network.

The aim of the following regression analyses was to determine if network structural properties of words significantly predicted performance in two language-related tasks. Predictors in the regression models included lexical variables (number of letters, number of phonemes, number of syllables, log of word frequency) and network variables (degree, clustering coefficient, closeness centrality). The dependent variables were Reaction Time (RT) and Accuracy (ACC) measures for two language tasks obtained from the English Lexicon Project—*speeded naming* (where participants read out loud the word shown on the screen) and *lexical decision* (where participants decide whether a presented letter string formed a word or not).

3 Results and Discussion

A two-step hierarchal regression was conducted with lexical variables added in Step 1 and network variables added in Step 2. In all models, the inclusion of network variables in Step 2 significantly improved model fit (see last row of Table 1), indicating that network variables were able to account for a small but significant amount of additional variance, beyond that of traditional lexical variables. Table 1 below shows a summary of the regression models at the final step.

Two key findings will be highlighted. First, degree was a significant predictor of naming and lexical decision performance. High degree words were processed more quickly and accurately than low degree words—consistent with previous work showing a processing advantage for words with many orthographic neighbors[6]. Second, closeness centrality was a significant predictor of naming and lexical decision performance. High closeness centrality words were processed more slowly and less accurately than low closeness centrality words in naming, whereas high closeness centrality words were processed more quickly than low closeness centrality words in lexical decision. In lexical decision, words that are “close” to many words may appear to be more “word-like”, such that participants take a shorter time to decide if a letter string is a word. On the other hand, high closeness centrality words may experience greater competition from other words in the lexicon, such that it worsens performance in the naming task where one has to retrieve the representation of a specific word from long-term memory.

Table 1. Summary of regression analyses of ELP data.

	Speeded naming				Lexical decision			
	RT		ACC		RT		ACC	
Predictors	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
	Step 1: Lexical variables							
Number of letters	6.43	< .001	4.54	< .001	-8.20	< .001	16.7	< .001
Number of phonemes	-2.96	.003	10.9	< .001	-11.9	< .001	14.4	< .001
Number of syllables	8.09	< .001	-14.3	< .001	18.7	< .001	-13.6	< .001
Log frequency	-44.9	< .001	40.3	< .001	-86.8	< .001	70.9	< .001
	Step 2: Network variables							
Degree	-13.0	< .001	8.46	< .001	-7.83	< .001	7.58	< .001
Clustering coefficient	-1.85	$.06^+$	2.14	.03	1.80	.07	-1.48	.14
Closeness centrality	4.00	< .001	-5.76	< .001	-3.87	< .001	0.73	.46
	$\Delta R^2 = .012$, F (3, 10705) = 60.7, $p < .001$		$\Delta R^2 = .006$, F (3, 10705) = 29.1, $p < .001$		$\Delta R^2 = .006$, F (3, 10705) = 40.7, $p < .001$		$\Delta R^2 = .005$, F (3, 10705) = 25.1, $p < .001$	

4 Conclusion

An analysis of the orthographic forms obtained from a large database revealed that the LCC of the orthographic language network is scale-free and consisted of a small-world structure with a degree distribution that approximated a power law. Regression analyses further revealed that various network characteristics significantly predicted performance on speeded naming and lexical decision. These results have important implications for the field of psycholinguistics because they demonstrate how network science can be used to quantify the structure of the mental lexicon and further our understanding of the cognitive processes that underlie visual word recognition.

References

1. Aitchison, J. (2012). Words in the mind: An introduction to the mental lexicon. John Wiley & Sons.
2. Steyvers, M., & Tenenbaum, J. B. (2005). The large‐scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*(1), 41-78.
3. Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language and Hearing Research, 51*(2), 408-422.
4. Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science, 20*(6), 729-739.
5. Siew, C. S. Q., & Vitevitch, M.S. (2016). Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42(3),* 394-410.
6. Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review, 4*(4), 439-461.
7. Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.
8. Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971–979.

Informed system improvement: Utilization of network analysis to assess health systems.

Stephanie Bultema[1], Hadley Morrow[2], Amy Riffe[3], Stacy Wenzl[4], Alison Karl White[5]

[1] University of Colorado Denver, Center on Network Science, Denver, Colorado, USA
stephanie.bultema@ucdenver.edu

[2, 5] Better Health Together, Spokane, Washington, USA

[3, 4] Spokane Regional Health District, Data Center, Spokane, Washington, USA

1 Overview

This submission details an underutilized approach to understanding health systems in the United States (US). Under Subtitle A – Transforming the Health Care Delivery System in the Patient Protection and Affordable Care Act of 2010, states are required to demonstrate improvements in the quality, value, and implementation of their health systems [1]. According to the World Health Organization (WHO), "a health system consists of all organizations, people and actions whose primary interest is to promote, restore, or maintain health" [2]. However, no US entity claims responsibility for coordinating, managing, or inventorying actors and their relationships in existent health systems, which poses a significant challenge for informed health system transformation. This study used survey research to gather quantitative and qualitative data from 165 representatives of 95 organizations participating in Eastern Washington's health system, which was used to generate a relational dataset with 564 actors and 9,039 ties. The regional health system was analyzed using network analysis and partitioned by relationship type, organizational sector, and geographic area. The areas of greatest activity included: collaboration and referral relationships; the health and social sectors; and Spokane County and Stevens County. The health system was least active with: data exchange, professional education, and financial support relationships; the business sector; and Adams County, Lincoln County, and Tribal Nations. Findings can be used as a health system assessment to understand the region's health system structure, including strong and weak points; as a planning tool to guide system improvement efforts; and as a baseline for evaluation of the health system over time. This study contributes to the academic literature by detailing an empirical application of network science theories and methods to the transformation of health systems under the Affordable Care Act.

2 Methods

The network of organizations participating in the population health and social determinants of health systems in BHT's six-county region was bound using a three-phase snowball sampling approach. Organization representatives who had formally engaged with BHT in some way were invited to participate in BHT's Population and Social Determinants of Health Systems Survey through four primary means: First, through direct invitation from a BHT Community Linkage Mapping project team member. Second, through direct invitation from another organization representative. Third, by on-site distribution and collection of the paper survey at BHT meetings. Fourth, through general BHT communication, including the BHT blog and newsletter. The survey was fielded from November 10, 2016 through March 6, 2017. Upon survey close, 165 responses from individuals representing 112 organizations had been collected.

For data collection, a survey instrument was developed which drew on various sources to inform the structure and question design [3, 4]. The survey collected relational and categorical data. Organization representatives were asked to respond to the survey from an organizational, rather than individual, perspective. There was no limit to the number of individuals who could respond from a single organization.

For analysis, Stata v. 13 was used to convert relational data from a Snap WebHost text file to an Excel CSV edges table. SPSS v. 22 was used to conduct descriptive analysis. Microsoft Excel was used to develop charts of descriptive data. Gephi 0.9.1 was used to conduct network analysis and develop network maps. ArcGIS Online was used to develop web-based, interactive, geographic network maps [5]. An infographic describing findings was developed in Piktochart [6].

3 Results

There were 564 organizations identified as participants in BHT's regional population and social determinants of health system. The system had a modularity of 0.2 with 62 identified communities. Organizations were categorized by sector based on BHT stakeholder groups: social, health, public, education, or business. Nearly half of the organizations identified as participants in BHT's regional health system were in the social sector (44%, n=249), with the other half comprised of organizations in the health (19%, n=108), public (18%, n=103), and education (14%, n=78) sectors. Relatively few organizations from the business sector (5%, n=26) were reported to have participated in BHT's regional health system in 2016. Of all organizations, 508 were reported to have maintained 9,039 community linkages in 2016. Most linkages reported in BHT's regional health system were collaboration (56%, n=5,050) or referral (30%, n=2726). The remaining linkages were data exchange (6%, n=517), education (5%, n=473), and financial support

(3%, n=273). About 1 in 10 organizations (57) identified as participants in BHT's regional health system had no identified linkages. On average, each organization maintained 31 linkages with up to 16 other organizations. Although an organization would only need an average of one introduction to a new organization to be linked to any other organization within BHT's regional health system (average path length=2.3), it could take up to three introductions to make a personal connection with a previously unknown organization (network diameter=4). About 3% of all possible inter-organizational linkages were reported (graph density=0.03).

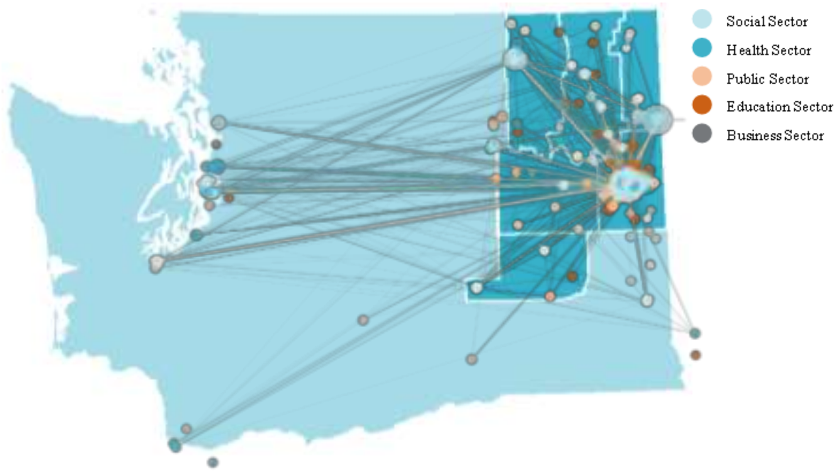


Figure 1 Eastern Washington's health system. Target counties are highlighted in teal. Each node represents a geocoded organization. Nodes are sized by average weighted in-degree and colored by sector. Edges are colored to match the sectors they link and are sized by weight. Nodes outside the state boundary represent organizations in other states (not shown by geocoded location).

4 References

[1] Patient Protection and Affordable Care Act, 42 U.S.C. § 3001. (2010).

[2] World Health Organization. (2017). *The WHO health systems framework.* Retrieved from http://www.wpro.who.int/health_services/health_systems_framework/en/

[3] Bultema, S. A. (2015). *Excelerate Success Pilot Baseline Network Analysis.* Retrieved from Excelerate Success website: https://goo.gl/DKKjkP.

[4] ReThink Health. (2015). *Baseline Network Assessments: Phase 1 Networks.* St. Louis, MO: Carothers, B.J., Sorg, A.A., Luke, D.A., Milstein, B.

[5] Spokane Regional Health District, Data Center. (2017). [Interactive geographic network maps of health systems by linkage type]. *Better Health Together Community Linkage Maps*. Retrieved from https://arcg.is/2pH9kuT

[6] Bultema, S. A. (2017). *Better Health Together Population and Social Determinants of Health System.* Retrieved from create.piktochart.com/output/23176905-web-bht-community-linkage-mapping

Producing official statistics from network data

Jan van der Laan[1] and Edwin de Jonge[1]

Statistics Netherlands (CBS), Henri Faasdreef 312 The Hague, the Netherlands,
dj.vanderlaan@cbs.nl
e.dejonge@cbs.nl

1 Introduction

The origin of Official Statistics is in the census, which is as old as civilization. Measuring sociodemographic properties, such as age, income and happiness, was, is and will be of importance for social scientists, historians, policy makers and government. A census is very valuable but its description of connections between people is very limited: it records people living on the same address, but it fails to capture the broader network of relationships that make up society: family, friends, neighbors, co-workers and acquaintances. Most demographic statistics describe (aggregates of) properties of inhabitants. From a perspective of measuring society a network can be a source of interesting demographic statistics. Is the strength of family ties regionally correlated? How diverse are personal networks? Given the current demographic trend in most countries that the average age is increasing, do parents live close to their children or is this distance increasing?

To address these kind of analyses and questions and to scan its potential for producing official demographical statistics a directed network with family, dwelling, neighbor, school going children and coworkers relationships was derived for the 17 million inhabitants of the Netherlands.

2 Network derivation: Who knows whom?

A complete and accurate derivation of a network that captures relationships between inhabitants would entail that all people that “know” each other are connected. Although many ingredients for constructing such a network are available, a complete acquaintance graph is in practice not possible. What can be done though, is to derive a network of people that are likely to know each other given auxiliary information. The resulting network resembles the real network, having similar network characteristics and statistics. The core of the network is formed by the population register which contains all persons registered in the Netherlands on October 1st 2014. The additional sources defining the edges are:

- The *parent-child register* is used to derive the following family relationships: ‘child-parent’, ‘sibling-sibling’, ‘grandchild-grandparent’, ‘nephew/niece-uncle/aunt’, ‘parent-parent’, ‘cousin-cousin’.
- The *household register* is used to derive the ‘household member’ relationship for all person living in the same dwelling.

- The location of each household is used to derive ‘neighbor’ relationships, for each person living in the ten closest households within 50 meters.
- The *employment register* is used to derive ‘co-workers’.
- The primary and secondary *education registers* are used to derive ‘children going to the same school’. Without more information and to constrain the number of edges only edges between persons of same age (in years) are used.

The resulting network contains 16.9 million vertices and 39.0 billion edges.

3 Preliminary results

A direct result of the derived network are family networks. Geographic distances of family members are of interest to policy makers because older persons are relying more and more on family care. Current research focuses on measuring ‘segregation’ in the Netherlands, which is suspected to increase: the working hypothesis is that network communities will be more homogeneous in income, education, ethnicity, neighborhood and schools, indicating segregation. We are planning to apply clustering methods to the network and investigate the dissimilarity between the clusters. Initially, we will be comparing networks of different regions.

Figure 1 shows network communities[1] in the Rotterdam area, in which only inhabitants of Rotterdam and their edges were selected. The Louvain method for community detection [3] using `igraph` R-package [4] was applied to this subgraph with 622 thousand vertices and 171 million edges. The figure displays a scatter plot matrix with ethnic group fractions, in which each dot identifies a community. The colors are determined using k-means clustering [1] on the ethnic fractions. It shows that not all groups are equally ‘mixed’. For example, communities with a large number of persons with a Turkish background (purple) have a small fraction of persons with a Moroccan background and vice versa for the clusters with a large number of persons with a Moroccan background (red). The mixing seems to be largest for persons with a Western, Other Non-Western and Antillean background. This is confirmed by the index of dissimilarity[2] calculated for these clusters (see Table 1).

Table 1. Index of Dissimilarity for each of the ethnic groups.

Ethnicity:	Native Dutch	Moroccan	Turkish	Surinam	Antilles	Other Non-West	Western
Dissimilarity:	0.305	0.439	0.435	0.394	0.219	0.202	0.119

4 Conclusion and discussion

The results presented are the first steps into investigating the possibilities using (derived) network data for official statistics purposes. The resulting in- and out-degree

[1] Communities containing less than 1 thousand persons were removed from the analysis.

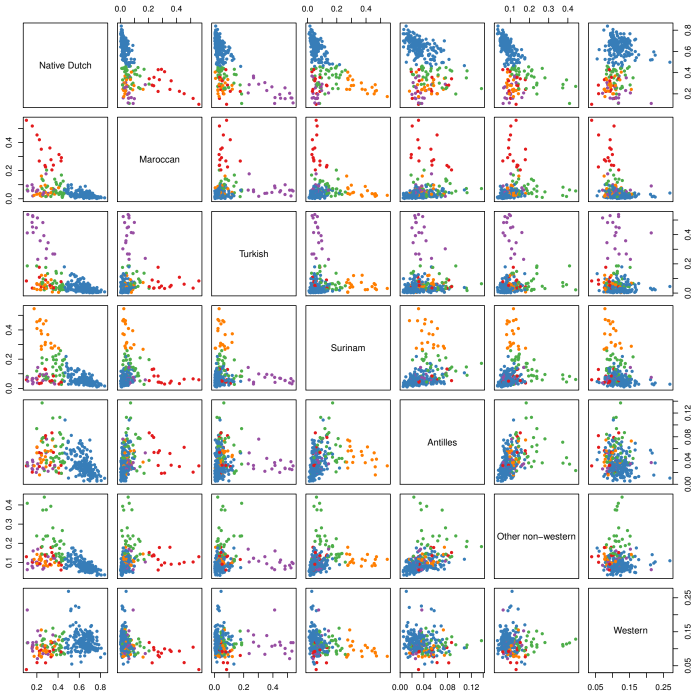


Fig. 1. Scatter plot matrix showing the fractions of each of the ethnicity groups in the communities detected in the graph. The colors are derived using k-means clustering.

distributions may be used to model networks of other countries and regions where summary statistics is available, but individual data is scarce. As for using the network to detect communities, further details need to be refined. For example, how does one weight the different types of relationships? How does one compare community structures of different regions/time periods? Current results do show that using graph analytics methods for analyzing society are promising.

References

1. Hartigan, J. A. and Wong, M. A.: A K-means clustering algorithm. Applied Statistics 28, 100-108 (1979).
2. Jahn, Julius, Calvin F. Schmid, and Clarence Schrag: The measurement of ecological segregation. American Sociological Review 12, no. 3 (1947): 293-303.
3. Blondel V.D., Vincent D., J.-L. Guillaume, R. Lambiotte and E. Lefebvre: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, (Oct 2008)
4. Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

Rich-clubness test: how to determine whether a complex network has or doesn't have a rich-club?

Alessandro Muscoloni[1] and Carlo Vittorio Cannistraci[1,2,*]

[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[2] Brain bio-inspired computation (BBC) lab, IRCCS Centro Neurolesi "Bonino Pulejo", Messina, Italy

Rich-clubness defines the extent to which a network is characterized by the presence of a cohort of nodes with a large number of links (rich nodes) that tend to be well connected between each other, creating a tight group (club) [1]. Although different methods, null-models and statistical tests have been proposed for assessing the rich-clubness, a procedure that assigns a unique value of rich-clubness significance to a given network is still missing.

Here we introduce three novelties. First, we designed the Cannistraci-Muscoloni (CM) null-model which is able to generate random networks with a lower rich-club coefficient, more suitable for the characterization of the rich-clubness. In particular, it was able to solve the main point of weakness of the Maslov-Sneppen (MS) procedure [2], which can produce random networks with a rich-club coefficient as high as in the original network, bringing to misleading conclusions (Fig. 1). Second, we introduced a new normalization of the rich-club coefficient using the difference rather than the ratio, favouring the detection of the correct peak of deviation from the null-model. Third, we proposed a new statistical test which is the first to assign a unique p-value to a given network. Furthermore, if the p-value is significant, the corresponding rich-club subnetwork is also detected. (Fig. 2). The test presents robustness since it exploits a population of random peaks of deviation from the average null-model that might occur also at different degree values. The advantage with respect to the state-of-the-art statistical test [3], [4], which provides a p-value for each degree, is that the latter one can produce significant p-values also for low-degree cohorts (paradoxically detecting a club that is not rich).

This study can have relevance for the analysis of the internal organization and function of networks arising in systems of disparate fields such as transportation, social, communication and neuroscience. In fact, simulations that investigate how the functional performance of a network is changing in relation to rich-clubness might be more easily tuned controlling one unique value that is the proposed rich-clubness measure.

References

[1] S. Zhou and R. J. Mondragón, "The Rich-Club Phenomenon in the Internet Topology," *IEEE Commun. Lett.*, vol. 8, no. 3, pp. 180–182, 2004.

[2] S. Maslov and K. Sneppen, "Specificity and Stability in Topology of Protein Networks," *Science*, vol. 296, no. 5569, pp. 910–913, 2002.

[3] Z. Q. Jiang and W. X. Zhou, "Statistical significance of the rich-club phenomenon in complex networks," *New J. Phys.*, vol. 10, 2008.

[4] M. P. van den Heuvel and O. Sporns, "Rich-club organization of the human connectome," *J. Neurosci.*, vol. 31, no. 44, pp. 15775–86, 2011.

[5] F. Papadopoulos, M. Kitsak, M. A. Serrano, M. Boguna, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, pp. 537–540, 2012.

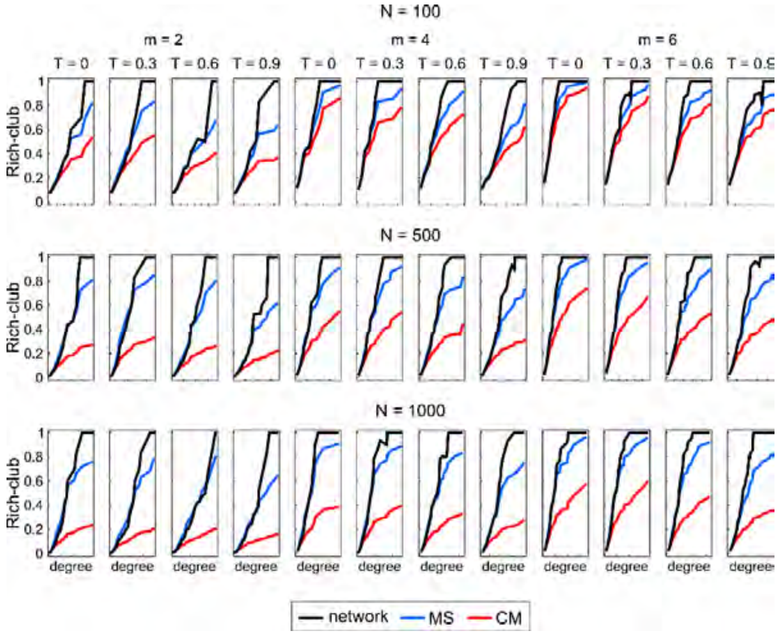


Fig. 1. We created a synthetic network for each combination of tuneable parameters of the Popularity-Similarity-Optimization (PSO) model [5] (size *N*, half of average degree *m*, temperature *T*), fixing the power-law degree distribution exponent $\gamma = 2.5$. For each PSO network we generated 1000 random networks using the two null-models discussed in this study: Maslov-Sneppen (MS) and Cannistraci-Muscoloni (CM). The plots report, for each different parameter combination, the rich-club coefficient (non-normalized) of the PSO network and the mean rich-club coefficient for each null-model, averaged over the 1000 repetitions. The figure highlights that for most of the parameter combinations, in particular for large network size (*N*) and average node degree (2*m*), the synthetic networks present fully connected *k*-subnetworks for high degrees, which might suggest the presence of rich-clubness. However, it can be noticed that also the MS null-model is characterized by a rich-club coefficient very close to the one of the original network. This represents a relevant limitation of the MS procedure when adopted in this context. Networks could be classified as non-rich-club not because the rich nodes do not form a club, but because the rich nodes form a club also in the null-model.

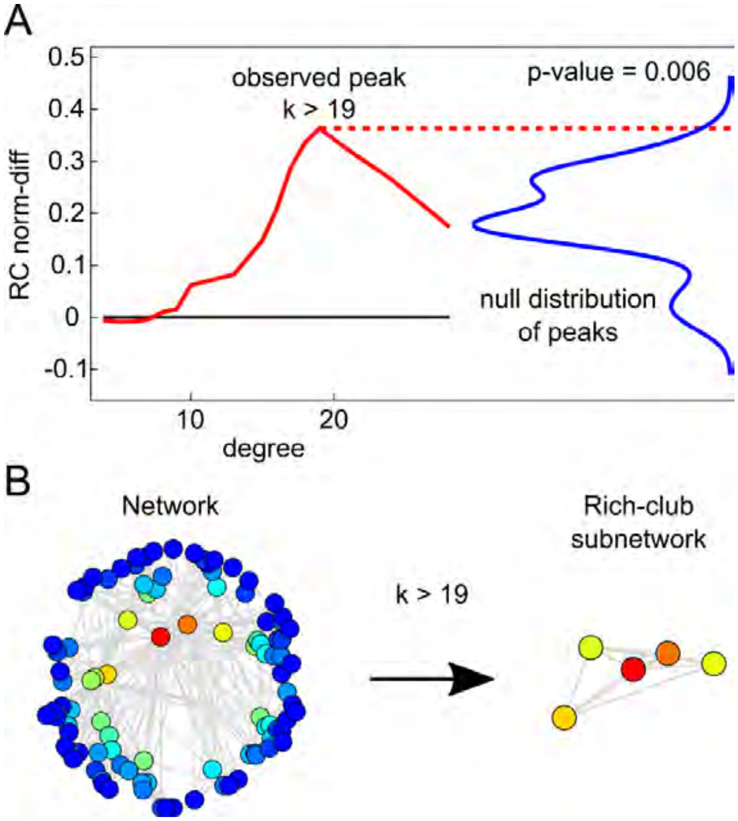


Fig. 2. (A) Explanatory plot of the statistical test for rich-clubness. (1) Given a network, a population of 1000 random networks are generated using the CM null-model proposed in this study. (2) The rich-club coefficient of the considered network is computed for each degree and then normalized (using the difference) by the mean coefficient of the random networks. The normalized coefficient is shown in red, whereas the black ground line indicates the reference case in which the rich-club coefficient is equal to the mean coefficient of the null-model. (3) The rich-club coefficient of every random network is also computed for each degree and normalized by the mean coefficient of the 1000 random networks. (4) The maximum value of the normalized rich-club coefficient (peak of deviation from the mean null-model) is computed both for the considered network (observed peak) and for the population of random networks (null distribution of peaks, shown in blue). Notice that to preserve clarity the 1000 normalized RC curves for the random networks are not reported. However, the blue line on the right shows the null distribution of random peaks generated by the random networks obtained by the CM null-model. The dashed red line represents the projection of the observed peak to the null distribution of random peaks. (5) A one-sided p-value is computed as the percentage of random peaks greater or equal than the observed peak. (B) The considered network is shown with nodes coloured by increasing log-degree using a cold-to-warm colour map. The rich-club subnetwork is highlighted, composed by the nodes with degree larger than the degree of the observed peak (k > 19) in the normalized rich-club coefficient.

Network reduction towards a scale-free structure preserving physical properties

Nicolas Martin, Paolo Frasca, and Carlos Canudas-de-Wit

Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP[†], GIPSA-Lab, 38000 Grenoble, France
{nicolas.martin, paolo.frasca, carlos.canudas-de-wit}@gipsa-lab.fr
www.scale-freeback.eu

1 Introduction

The analysis and control of dynamics on large networks is often a challenging task. Reduction methods which can cut this complexity have intensively been developed and see for example these recent works [1, 2]. We propose here a new approach of reduction allowing to drive an original large arbitrary network towards a particular structure: a scale-free network. Scale-free networks are characterized by a large number of nodes with a small degree and some nodes with a very large degree: their degree distribution is a power law. The scale-freeness of a network implies interesting properties: robustness to failures, ultra-small world property [3], and interesting features for control design [4]. Then, it appears that, for some applications, the reduction method would be more interesting if it allows to obtain a reduced network with a scale-free structure. The objective of the method developed here is starting from an arbitrary large weighted and directed network and finding a reduced network with a scale-free structure, while preserving some physical properties of the initial graph for consistency.

2 Problem formulation

The reduction is made by partitioning the initial graph into groups of connected nodes and assigning each part as a new node in the reduced graph. By this way, the topological structure of the graph (interconnections) is preserved. Then the problem can be viewed as a minimization problem where we look for a partition of the initial graph G_0 such that the graph $\tilde{G}$ coming out of this partition minimizes the sum of two cost functions: $\mathrm{J}_{\mathrm{SF}_\alpha}(G) + \mathrm{J}_{\mathrm{sim}}(G, G_0)$ where $\mathrm{J}_{\mathrm{SF}_\alpha}(G)$ is the scale-free error and gives indication on the scale-freeness of the graph and $\mathrm{J}_{\mathrm{sim}}(G, G_0)$ is the similarity cost function between G and G_0 and gives an indication on the error between the graph and the initial graph. The minimization is done under a constraint: $\tilde{G}$ has to preserve some physical properties.
The general problem being stated, we give some specifications to solve a particular case:

- The scale-free error is the 2-norm difference between the degree distribution of the graph and the power law corresponding to the desired scale-free distribution.

[†] Institute of Engineering Univ. Grenoble Alpes

- The similarity cost function is written as follow:

$$\mathrm{J}_{\mathrm{sim}}(G_1, G_0) = \frac{\|x^\star_{G_1} - \mathrm{Proj}(x^\star_{G_0})\|_2}{\|\,\mathrm{Proj}(x^\star_{G_0})\|_2} \quad (1)$$

  where $x^\star_G$ is an indicator of spectral centrality in the graph G (not detailed here), and $\mathrm{Proj}(x)$ is the vector of the sum of the element of x within each clusters. This cost function translates the closeness between the centrality in the reduced graph and the sum of the centrality within each cluster in the initial graph.
- The only physical property we want to preserve is the mass conservation: for every node the sum of the weights of incoming edges is equal to the sum of the weights of outgoing edges.

3 Results: Theorem and simulations

Definition: A merging is a particular partition where only two connected nodes are gathered. Precisely if the set of edges is $V = \{1,...,n\}$, a merging is a partition with the form $\{\{1\},...,\{v-1\},\{v+1\},...,\{w-1\},\{w+1\},...,\{n\},\{v,w\}\}$ where (v,w) is an edge of the graph.

Within the specifications given previously we have the following theorem.

Theorem 1. *Let G_0 be a weighted graph. For all merging M, the weights of the graph G_1 issued of the merging M of G_0 can be chosen such that G_1 preserves the mass conservation property and such that the similarity cost function is null.*

By exploiting this theorem, we have designed an iterative algorithm to solve the above optimisation problem: starting from an initial large graph, we look at each step for the best merging which is the merging whose resulting graph minimizes the scale-free error. We do not look at the value of the similarity cost function as it is null for all merging just by choosing the good weights. Instead of looking for the best edge within the whole set of edges, we only search within a small randomly selected set of edges. This choice allows to cut considerably the computation time and figure 2 shows that it has a minor influence on the efficiency of the algorithm. The complexity of the algorithm is $O(N_r N_v^2)$, where N_v is the number of nodes in the initial graph and N_r the size of the random set of edges tested.

Although the algorithm does not provide the exact solution of the problem, simulations show that the graph obtained is really close to the scale-free structure desired. Figure 1 shows the results of the simulation on the urban traffic network of Grenoble which is originally rather far from a scale-free structure.

Acknowledgements

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement N 694209)

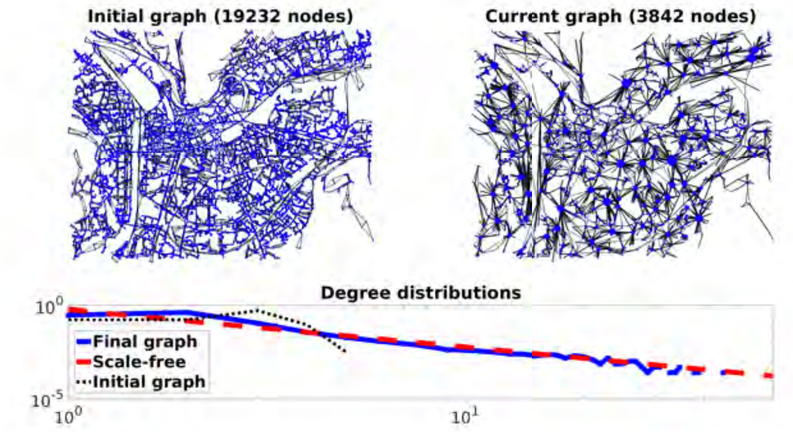


Fig. 1: Top left: the initial graph of Grenoble (double-way roads are thicker than one-way ones). Top right: result of the algorithm, a scale-free graph five times smaller than the initial one. The graph respects the mass conservation property and allows to reconstruct the centrality of each zone with a perfect consistency with the centrality in the initial graph. Bottom: the degree distributions show the ability of the algorithm to drive the initial graph to the desired structure.

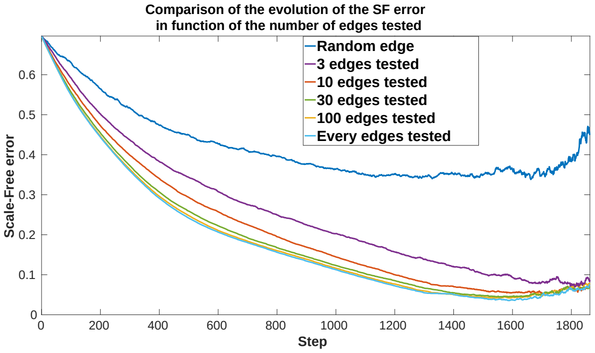


Fig. 2: Evolution of the scale-free error through the algorithm for different values of the size of the random set of edges: there is no significant advantage to explore a very large set of edges.

References

1. Ishizaki, T., Kashima, K., Girard, A., Imura, J. I., Chen, L., Aihara, K. (2015). Clustered model reduction of positive directed networks. Automatica, 59, 238-247.
2. Cheng, X., Kawano, Y., Scherpen, J. M. (2016). Graph structure-preserving model reduction of linear network systems. In Control Conference (ECC), 2016 European (1970-1975). IEEE.
3. Newman, M. E. (2003). The structure and function of complex networks. SIAM review, 45(2), 167-256.
4. Nacher, J. C., Akutsu, T. (2012). Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control. New Journal of Physics, 14(7), 073005.

Ranking of World Universities from 2017 Wikipedia Network

Célestin Coquidé[1], José Lages[1], and Dima L. Shepelyansky[2]

[1] Institut UTINAM, Observatoire des Sciences de l'Univers THETA, CNRS, Université de Bourgogne-Franche-Comté, Besançon 25030, France
http://perso.utinam.cnrs.fr/~lages/

[2] Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France
http://www.quantware.ups-tlse.fr/dima/

1 Introduction

The efficiency of academic education is nowadays a matter of political, economical and societal importance. University rankings, reviewed e.g. in many details in [4], are among the most important tools to measure academic efficiency. The available ranking approaches are based on human selection rules which can not be exhaustive or can favor certain cultural choices and preferences. Thus it would be highly desirable to construct an independent mathematical statistical method which ranks universities independently of any human rules. In order to fill this gap, we have recently proposed the Wikipedia Ranking of World Universities (WRWU) [5,7] based on the statistical analysis of networks of Wikipedia articles. Wikipedia now supersedes Encyclopedia Britannica in size and even in accuracy of articles devoted to many scientific domains [3]. Presently, Wikipedia contains more than 280 language editions representing different and complementary cultural views on human knowledge. This huge amount of encyclopedic data encodes also hidden information about how different cultures and societies are entangled. For all these reasons probing Wikipedia is relevant to build rankings for various aspects of human activities, one of these being higher education.

2 Methods

The mathematical grounds of this approach are based on Markov chain theory and, in particular, on the Google matrix analysis initially introduced in 1998 by Google's co-founders, Brin and Page [1], for hypertext analysis of the World Wide Web. Let's consider the network of the N articles of a given Wikipedia edition. The network adjacency matrix element A_{ij} is equal to 1 if article j points towards article i and equal to zero otherwise. The Google matrix element $G_{ij} = S_{ij} + (1-\alpha)/N$ gives the transition probability that a random reader jump from article j to article i. The stochastic matrix element is $S_{ij} = A_{ij}/\sum_{i=1}^{N} A_{ij}$ if article j is not a dangling node, otherwise $S_{ij} = 1/N$. The dumping factor $\alpha = 0.85$ allows the random reader to escape from dangling subnetworks. The right eigenvector $\mathbf{P}$ corresponding to the $\lambda = 1$ Google matrix eigenvalue is the PageRank vector. The vector element P_i is proportional to the number of times the random reader reads article i. The CheiRank vector $\mathbf{P}^*$ is the $\lambda = 1$ right eigenvector of the Google matrix constructed with the inverted network using A_{ji} instead of A_{ij}. PageRank measures the relative influence of nodes. Recursively, the more a node is pointed by influent nodes, the more it is influent. CheiRank measures the relative communicative ability of nodes. Recursively, the more a node points toward important communicative nodes, the more it is communicative. The ranking of the most influent (communicative) universities is obtained by extraction from PageRank (CheiRank) the articles devoted to universities.

3 Results

Table 1 (leftmost column) gives for the 2017 English edition of Wikipedia, the top10 of the most influent universities using PageRank algorithm. As a comparison, the top10 of the 2017 Academic Ranking of World Universities (ARWU) is shown in Table 1 (rightmost column). The two top10s (top100s; not shown) have 9 (61) universities in common confirming the fact that Wikipedia ranking is indeed able to measure academic excellence. Comparing each of these two rankings with the corresponding ones in 2013, we see that Wikipedia ranking is more robust since 9 universities are in common and keep their positions [7] and for ARWU 10 universities are in common but only 4 keep their positions. Fig. 1 (left panel) shows the geographical distribution of top100 universities from the 2017 English Wikipedia network PageRank analysis. As in ARWU, Anglo-Saxon universities dominate in number. Fig. 1 (right panel) gives the distribution of the density of 2017 English Wikipedia articles in the plane of PageRank index K *vs.* CheiRank index K^*. We clearly see that the most influent universities (low PageRank index K) are also among the top100 of the most communicative universities (CheiRank K^*); PageRanking and CheiRanking share 41% universities in common. We observe that top influent universities have a PageRank as low as $\sim 10^2$ indicating the very importance of the corresponding articles in the 2017 English Wikipedia network ($\sim 5 \times 10^6$ articles). Also, most of the ARWU top100 universities which are not present in Wikipedia top100 PageRank are also not present in the top100 CheiRank, indicating their lack of communicative ability via Wikipedia.

2017 English Wikipedia PageRanking	Rank	2017 ARWU
Harvard University	1	Harvard University
University of Oxford	2	Stanford University
University of Cambridge	3	University of Cambridge
Columbia University	4	MIT
Yale University	5	University of California, Berkeley
Stanford University	6	Princeton University
MIT	7	University of Oxford
University of California, Berkeley	8	Columbia University
Princeton University	9	California Institute of Technology
University of Chicago	10	University of Chicago

Table 1. Comparison between the ranking of world universities obtained from the PageRank of the 2017 English Wikipedia network (leftmost column) and the 2017 academic ranking of world universities provided by the Shanghai Jiao Tong University (rightmost column).

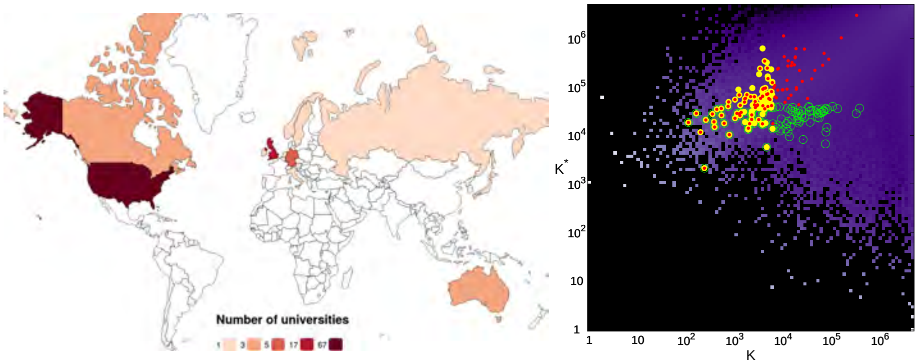


Fig. 1. Left panel: geographical distribution of the top100 universities obtained from the PageRanking of the 2017 English Wikipedia network. Right panel: density distribution $dN/dKdK^*$ of 2017 English Wikipedia articles in the plane of PageRank and CheiRank indexes (K,K^*) shown by color with dark violet for minimum and white for maximum (black for zero). Yellow disks (green circles) indicate the top100 universities using PageRank (CheiRank) algorithm. Red points indicate 2017 ARWU top100.

Fig. 1 shows results for 2017 English edition of Wikipedia. At the Complex Networks 2017 conference, we will present an exhaustive study of 24 different language editions of Wikipedia ($\sim 17 \times 10^6$ articles) representing about 60% of the total articles in Wikipedia and corresponding to about 60% of the total world population. Consequently, we will construct a network of culture comparing the different cultural point of views encoded in these language editions. Aggregating rankings for the 24 Wikipedia editions, we will provide the 2017 global Wikipedia Ranking of World Universities. Also, using the recently developed reduced Google matrix method [2,6], we will present hidden links existing between the most influent universities.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 30(1-7), 107–117 (Apr 1998), http://dx.doi.org/10.1016/S0169-7552(98)00110-X
2. Frahm, K.M., Jaffrès-Runser, K., Shepelyansky, D.L.: Wikipedia mining of hidden links between politicalleaders. The European Physical Journal B 89(12), 269 (Dec 2016), https://doi.org/10.1140/epjb/e2016-70526-3
3. Giles, J.: Internet encyclopaedias go head to head. Nature 438, 900–901 (Dec 2005)
4. Hazelkorn, E.: Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence. Palgrave Macmillan (2015)
5. Lages, J., Patt, A., Shepelyansky, D.L.: Wikipedia ranking of world universities. The European Physical Journal B 89(3), 69 (Mar 2016), https://doi.org/10.1140/epjb/e2016-60922-0
6. Lages, J., Shepelyansky, D., Zinovyev, A.: Inferring hidden causal relations between pathway members using reduced google matrix of directed biological networks. bioRxiv (2017), https://www.biorxiv.org/content/early/2017/02/06/096362
7. Lages, J., Shepelyansky, D.: WRWU website. http://perso.utinam.cnrs.fr/~lages/datasets/WRWU/ ([Online; accessed 06-October-2017])

Spatial network investigation of wall turbulence

Giovanni Iacobello[1], Stefania Scarsoglio[1], Hans Kuerten[2], and Luca Ridolfi[3]

[1] Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin, Italy
[2] Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands,
[3] Department of Environmental, Land and Infrastructure Engineering, Politecnico di Torino, Turin, Italy
giovanni.iacobello@polito.it

1 Introduction

In the last decades, complex networks have been exploited in a wide range of applications, with an increasing interest to physical and engineering problems. Specifically, the study of problems related to fluid flows mainly concerns two-phase flows [1] and geophysical flows [2], with particular attention to time-series mapping into networks (e.g., visibility [3]). In this work, we propose a correlation network-based investigation [4–7] of a turbulent channel flow, with the aim to spatially characterize the flow dynamics [8], introducing a novel statistical approach to wall turbulence analysis.

2 Database Description and Methods

The data [9] were extracted from a direct numerical simulation of a fully developed turbulent channel flow at $Re_\tau = Hu_\tau/\nu = 180$, where $H = 1$ is the half-channel height, $\nu = 1/180$ is the kinematic viscosity and $u_\tau = 1$ is the friction velocity. The physical domain is $(4\pi H \times 2H \times 4/3\pi H)$ with a grid resolution of $(N_x \times N_y \times N_z) = (576 \times 191 \times 288)$, where (x, y, z) are the streamwise, wall-normal and spanwise directions, respectively. The velocity field was computed and data were acquired for 5000 time samples, with a time-step $\Delta t = 2.510^{-4}$. Firstly, we selected $N'_x = 144$ equally spaced grid points in the x direction and $N'_z = 150$ consecutive grid points in the $z-$direction. We then assigned a node to each selected grid point, resulting in $n = (144 \times 191 \times 150) = 4125600$ nodes. The Pearson correlation coefficients based on the streamwise velocity component time-series were evaluated for each pair of nodes. Although such procedure can also be carried out for other physical quantities, the streamwise velocity is one of the most significant variables to characterize turbulent channel flows. A spatial network was then built, where links are active if the absolute value of the correlation coefficient is greater than or equal to a suitable threshold, τ, which was here set equal to 0.85. A high value of τ was chosen to highlight the strongest spatial correlations and to have a manageable number of links. In order to take into account the non-uniform spacing of the grid in the wall-normal direction, we assigned to each node a weight equal to the volume, V_j, of that node. Specifically, we defined the *volume-weighted connectivity* [10] of node i as $VWC(i) = \sum_j^N a_{ij} V_j / V_{tot}$, where a_{ij} are the entries of the adjacency matrix (with $a_{ii} = 1$), and V_{tot} is the total volume of the domain. $VWC(i)$ ranges between zero and one, and represents the fraction of volume to which the node i is connected.

3 Results

In Fig. 1(a) we highlight the nodes with high VWC values; in particular, we say that a node has a high VWC value if its cumulative probability, defined as $P(VWC) = \sum_{VWC'=VWC}^{\infty} p(VWC')$, satisfies $P(VWC) \leq 10^{-2}$ (corresponding to the 99th percentile), where $p(VWC)$ is the VWC probability distribution [11]. The nodes connected to a higher fraction of volume tend to group into clusters elongated in the streamwise direction, x, representing regions of high kinematic coherence, i.e. nodes with stronger spatial connections. Such clusters of *hubs*, as also evidenced in Fig. 1(b), are present both close to the walls (i.e., at $y^+ = 0, 360$) and at the center of the channel (i.e., at $y^+ = 180$), although the dynamics leading to the connections change at different wall-normal distances. It is interesting in Fig. 1(b), the nesting behavior of high VWC nodes.

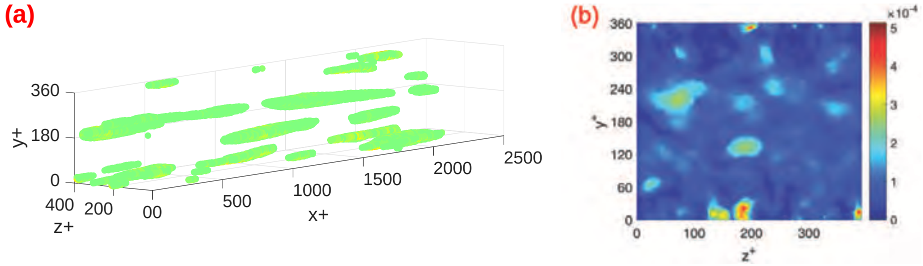


Fig. 1. *(a)* View of nodes with VWC in the 99th percentile. *(b)* Planar section at $x^+ = 2000$; the colorbar refers to the full range of VWC values. The axes are reported in wall-units, i.e. $(x^+, y^+, z^+) = (x, y, z) \cdot u_\tau / \nu$.

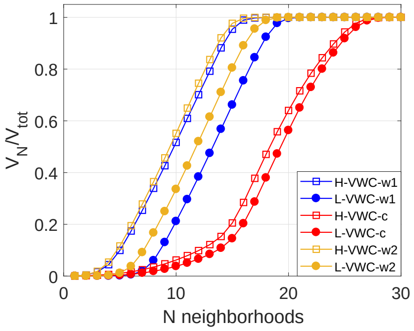


Fig. 2. The fraction of volume of the first N neighborhoods for three pairs of source-nodes. Notation: H/L-VWC, high/low VWC values; w1/w2, walls 1 and 2; c, center of the channel.

To have a better spatial characterization, the location of progressive neighbors of nodes at different VWC is investigated next. By doing so, it is possible to understand how different regions in the channel are connected. To this end, we analyzed the *successive neighborhoods* [12] of selected *source-nodes*, at different wall-normal distances.

We considered the N *cumulative neighborhoods* as the union of the first N neighborhoods. In order to illustrate the differences in the progressive patterns of the cumulative neighborhoods for different source-nodes, we selected three pairs of representative nodes — two pairs close to the walls and one at the center —, each one with a high and a low VWC value. Fig. 2 shows the behavior of the the fraction of the total volume occupied by the nodes in the first N neighborhoods for the three pairs of nodes. Starting from the source-nodes close to the wall, the expansion of the neighborhoods is clearly much faster than the expansion of the neighborhoods of source-nodes at the center of the channel, independently of the VWC values. However, at a fixed y^+, the expansion of the neighborhoods is slower for nodes with low VWC. In this context, an important role is played by nodes connected with negative correlation links, since they are generally less frequent and could suggest the presence of particular kinematic dynamics.

Summary. The present analysis can highlight spatial relations among different regions in a novel view that relies on the network formalism. In particular, the presence of clustered and elongated groups of highly connected nodes at various wall-normal distances indicates a spatial coherence related to the streamwise velocity correlations. The analysis of progressive neighborhoods, feasible only through a network approach, offers useful information on the spatial propagation of high-correlation patterns. Therefore, the proposed approach can provide new insights into the spatial characterization of complex systems such as turbulent flows, which deserves additional extensive investigation.

References

1. Z.K. Gao, W.X. Wang, N.D. Jin: Nonlinear Analysis of Gas-Water/Oil-Water Two-Phase Flow in Complex Networks, Springer Press (2014)
2. Ser-Giacomi, E., Rossi, V., Lpez, C., Hernndez-Garca, E.: Flow networks: A characterization of geophysical fluid transport. Chaos, 25(3), 036404 (2015)
3. Lacasa, L., Luque, B., Ballesteros, F., Luque, J., Nuno, J. C.: From time series to complex networks: The visibility graph. P. Natl. Acad. Sci., 105(13), 4972-4975 (2008)
4. Massara, G. P., Di Matteo, T., Aste, T.: Network filtering for big data: triangulated maximally filtered graph. J. of complex Networks, 5(2), 161-178 (2016)
5. Mantegna, R. N.: Hierarchical structure in financial markets. Eur. Phys. J. B, 11(1), 193-197 (1999)
6. Tumminello, M., Aste, T., Di Matteo, T., Mantegna, R. N.: A tool for filtering information in complex systems. P. Natl. Acad. Sci. USA, 102(30), 10421-10426 (2005)
7. Aste, T., Di Matteo, T.: Dynamical networks from correlations. Physica A, 370(1), 156-161 (2006)
8. Scarsoglio, S., Iacobello, G., Ridolfi, L.: Complex networks unveiling spatial patterns in turbulence. Int. J. Bifurcat. Chaos, 26(13), 1650223 (2016)
9. Computational resources were provided by HPC@POLITO (http://www.hpc.polito.it) and SURFsara-Cartesius (https://userinfo.surfsara.nl/)
10. Donges, J. F., Zou, Y., Marwan, N., Kurths, J.: Complex networks in climate dynamics. Eur. Phys. J.Spec. Top. 174(1),157–179, (2009)
11. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D. U.:Complex networks: Structure and dynamics. Phys. Rep., 424(4), 175-308(2006)
12. McAuley, J. J., da Fontoura Costa, L., Caetano, T. S.: Rich-club phenomenon across complex network hierarchies. Appl. Phys. Lett., 91(8), 084103 (2007)

Degree in Probabilistic Networks: Revisited

Amin Kaveh, Matteo Magnani, and Christian Rohner

InfoLab, Department of Information Technology, Uppsala University, Sweden
amin.kaveh@it.uu.se,
matteo.magnani@it.uu.se,
christian.rohner@it.uu.se

1 Introduction

Probabilistic networks where probabilities are assigned to the edges can be used to model systems where the existence of connections between its entities cannot be determined with certainty. For example, more than 50% of the edges in protein-protein interaction networks do not correspond to actual interactions [4], edges in computer communication networks can be subject to failures, and edges in online social networks may not correspond to offline social interactions.

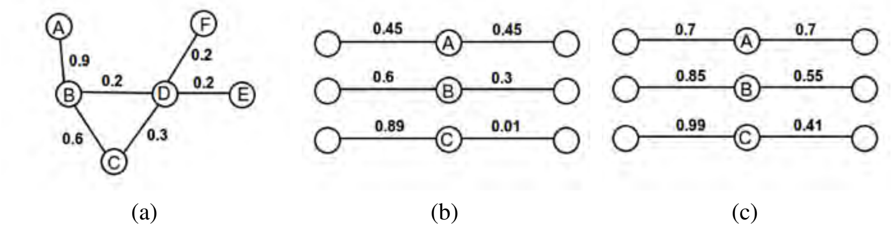
The most prevalent measure corresponding to node degree in probabilistic networks is *expected degree* which is the sum of the probabilities of the edges connecting to the node [3] [2]. In this work we show that, despite its popularity as a centrality measure for probabilistic networks, expected degree can produce counter-intuitive results. We then define basic properties that degree centrality measures in probabilistic networks should satisfy, showing that expected degree and related measures do not satisfy them. We conclude by discussing alternative approaches.

2 Limitations of existing measures

In the following we will notate $P_>(i,j)$ the probability that the degree of node i is higher than the degree of node j — that is, $P(deg_i > deg_j)$, and $P_=(i,j)$ the probability that the degree of node i equals the degree of node j — that is, $P(deg_i = deg_j)$.

Fig. 1(a) shows a probabilistic graph where expected degree does not succeed in providing a satisfactory representation of node connectivity. As an example, nodes A and D have the same expected degree (0.9). However, D's degree can be up to four times higher than A. Conversely, assuming probabilistic independence between edges, $P_>(A,D) > P_>(D,A)$. Expected degree does not capture any of these features. A similar reasoning can be applied to other combinations of nodes.

While Fig. 1(a) shows that not only the total probability but also the number of edges plays a role in characterizing node centrality, even these two statistics together are still not enough to provide a good characterization of node degree in probabilistic networks. For the three nodes in Fig. 1(b), and Fig. 1(c), Table 1 indicates both the expected degree and the generalized degree centrality [1], a degree measure for weighted networks considering also the number of edges. It can be noticed how the three nodes are indistinguishable according to these two measures. However, in Fig. 1(b) $P_>(A,C) < P_>(C,A)$, while in Fig. 1(c) $P_>(A,C) > P_>(C,A)$. Similar considerations apply to other nodes in the figures.



(a) (b) (c)

Fig. 1. (a) A probabilistic Network, where nodes A, C and D have the same expected degree. (b-c) Different distributions of probabilities across the edges of three nodes.

Table 1. Comparing the nodes in Figures 1(b) and 1(c). *E*: Expected Degree, *GD*: Generalized Degree. The three nodes in each figure are indistinguishable using these statistics, even if for some nodes there is a higher probability that its degree is higher than the others

Graph	node	E	GD	Graph	node	E	GD
Fig 1(b)	A	0.9	$2^{\alpha}\, 0.9^{1-\alpha}$	Fig 1(c)	A	1.4	$2^{\alpha}\, 1.4^{1-\alpha}$
	B	0.9	$2^{\alpha}\, 0.9^{1-\alpha}$		B	1.4	$2^{\alpha}\, 1.4^{1-\alpha}$
	C	0.9	$2^{\alpha}\, 0.9^{1-\alpha}$		C	1.4	$2^{\alpha}\, 1.4^{1-\alpha}$

3 Degree in probabilistic networks: basic properties

In the previous section we have given an intuition of the fact that expected degree does not capture well the relationship between node degrees in probabilistic networks. A degree function $f(i)$ that can capture node degrees in probabilistic networks, would ideally fulfill the following two properties. $\forall i, j$:

$$P_{>}(i,j) = P_{>}(j,i) \Rightarrow f(i) = f(j) \tag{1}$$

which means that regardless of the value of the probability that degree of nodes i and j is equal, if $P_{>}(i,j) = P_{>}(j,i)$ then their degree have to be equal, and:

$$P_{>}(i,j) > P_{>}(j,i) \Rightarrow f(i) \geq f(j) \tag{2}$$

Where, $P_{>}(i,j) + P_{>}(j,i) + P_{=}(i,j) = 1$.

4 Towards new degree centrality measures

In our presentation, we will discuss alternative ways to define a degree centrality measure satisfying the previous properties. The main idea behind our discussion is that degree centrality in probabilistic should be a function of (1) the total (or average) probability mass assigned to edges adjacent to the node, (2) the number of edges and (3) the distribution of the probability mass across the edges.

A highlight of our results is presented in Fig. 2, where we compare two nodes with extreme patterns of probability mass distribution across their edges in the case of nodes with two adjacent edges (this can be generalized to arbitrary degrees). In particular, one node has its probability mass equally distributed across its edges and the second has most of the probability mass associated to a single edge. As we can see in the figure, the average edge probability (E/l) gives us an indication of which node has a higher probability of having a higher degree. In particular, when $E/l = 0.5$, regardless of how probabilities are distributed across edges $P_>(i,j) = P_>(j,i)$. If $0 < E/l < 0.5$, $P_>(i,j) < P_>(j,i)$ where i is the node with equally distributed probabilities. On the other hand, if $1 > E/l > 0.5$, $P_>(i,j) > P_>(j,i)$.

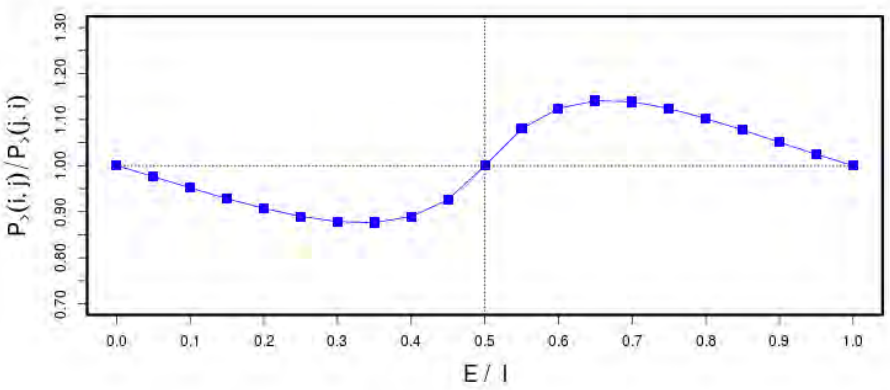


Fig. 2. $l_i = l_j = 2$, $E_i = E_j$.

References

1. Opsahl, T., Agneessens, F., Skvoretz, J.: Node centrality in weighted networks: Generalizing degree and shortest paths. Social networks 32(3), 245–251 (2010)
2. Parchas, P., Gullo, F., Papadias, D., Bonchi, F.: Uncertain graph processing through representative instances. ACM Transactions on Database Systems (TODS) 40(3), 20 (2015)
3. Parchas, P., Gullo, F., Papadias, D., Bonchi, F.: The pursuit of a good possible world: extracting representative instances of uncertain graphs. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data. pp. 967–978. ACM (2014)
4. Von Mering, C., Krause, R., Snel, B., Cornell, M., et al.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887), 399 (2002)

Concepts co-occurrence for the identification of higher order concepts in Mathematics articles.

Vsevolod Salnikov[1], Renaud Lambiotte[1,2], and Daniele Cassese[1]

[1] University of Namur, Namur 5000, Belgium,
[2] Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK,
vsevolod.salnikov@unamur.be
renaud.lambiotte@maths.ox.ac.uk
daniele.cassese@unamur.be

1 Introduction

Words co-occurrence networks capture relationships between words co-occurring in a given document or phrase: each node is a word, and an edge exists if two words both appear in the same document. Co-occurrences networks have been used, among other things, to study the structure of human language [5], to detect influencial text segments [7], to identify authorship signature in temporal evolving networks [1]. Here we are interested in the analysis of the mathematical content of scientific articles, in order to develop measures for visualising mathematical concepts that can be used to classify scientific production in a novel way. The use of co-occurrence networks in scientometrics is not new, for example [10] focus on weighted co-occurrences networks of keywords to study the expansion of knowledge in nanoEHS risk literature.

Our work differs as we extend our analysis a large set of mathematical concepts, and because we use a simplicial complex framework, as in [4], to have a better understanding of the clustering of concepts within and across different articles. Modeling co-occurrence relations as a simplicial complex allows to go beyond the network description that reduces all the structural properties to pairwise interaction and their combinations, and provides a natural framework for studying higher-order relations. This approach has proved to be very useful when the data has inherently rich structure, as in [9].

2 Dataset

In this work we analyse the dataset of 54K arXiv articles, divided into subject subsets, extracting from each article text its mathematical content. We extract mathematical concepts from Wikipedia, including 1618 expressions, and all the equations in LaTeX format.

The resulting network is very dense, hence exploratory analysis of communities with Louvain [3] method is not very informative, as the compression of data induced by the network approach inevitably causes loss of information about the conceptual layers. This is a common problem with co-occurrences networks, we decided not to address it thresholding edges below or above a certain weight, as we want to keep all the information, as we cannot rule out that some low-weight connections reveal important bridges between conceptually different areas. Instead we use the Persistent Homology method [8] to explore the higher dimensional shape of our data.

3 Results and Discussion

The concept co-occurrences network provides a low-level visualisation of the conceptual content of the articles, and a framework to analyse the similarities between the content of articles, that can be used to develop a much finer classification of scientific production than that based on journals or keyword co-occurrences networks and allows to explore the effective distance between different fields on the basis of an objective metric. Moreover, studying the temporal variability of both the co-occurrences and articles networks we aim to draw a picture of the evolution of scientific disciplines, and the emergency of new fields as the merging of previously separated areas.

Loosely on the line of Russell's and later Wittgenstein's logical atomism, we consider the nodes of our network as the atoms of mathematical concepts which relations are captured by their co-occurrence in a scientifically coherent document. Conceptual atoms together with their relations originate conceptual complexes, that we visualise as aggregations of conceptual simplices: thus, detecting higher level concepts in our dataset is equivalent to find higher order structures on the co-occurrence network, and study their relations.

Using a weight rank clique filtration [9, 8] on the weighted co-occurrences simplicial complex, we analyse its persistent homology H_p, that detects how the pth homology changes along the steps of the filtration. H_0 detects connected components, H_1 loops, H_2 voids (or 2-dimensional holes). One of the characteristics that immediately emerges from our dataset is that there are many persistent holes of different dimensions in our dataset. In H_1 there are 11972 loops, of which only 7 die before the final step of the filtration, while the analysis of H_2 yields 22770 voids, most of which 'live' till the end of the filtration, and only 8 'die' at intermediate stages. Voids are particularly interesting as, in representing lower connectivity between concepts, detect potential 'missing' triangles of co-occurrences that can unify different areas of mathematics. Understanding which is the minimum number of triangles needed is one of our questions, that we hope to answer with further research, in order to predict which are the missing connections in those homological holes that never die that would change the structural shape of mathematics.

References

1. Akimushkin, C., Amancio D.R., Oliveira O.N. Jr.: Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks. PLoS ONE. 12(1), 1 – 101 (2017)
2. Börner, K., Chen, C. and Boyack, K.: Visualizing knowledge domains, Annual Review of Information Science & Technology, 37 179–255 (2003)
3. Blondel V., Guillaume J.-L., Lambiotte R., Lefebvre E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment (10), P10008, (2008)

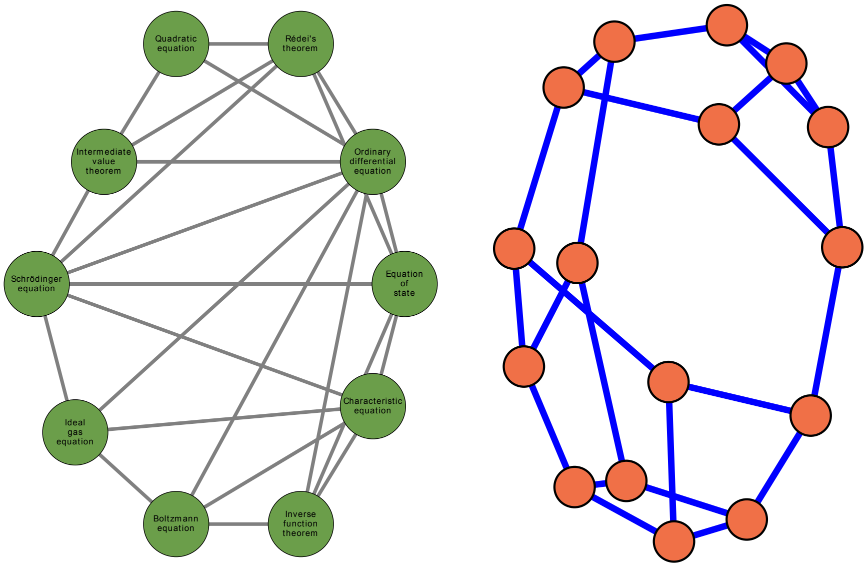


Fig. 1. The smallest (10-triangles) short-lived void in our dataset (left) and its corresponding factor graph (right). In the factor graph each node is a triangle of concepts, and there is an edge between two triangles each time they share a side

4. Chiang I-J.: Discover the semantic topology in high-dimensional data. Expert Systems with Applications. 33 256-262 (2007)
5. Ferrer i Cancho, R., Solé R.V.: The small world of human language. Proc. R. Soc. Lond. B 2001 268 2261-2265; DOI: 10.1098/rspb.2001.1800. (2001)
6. Ferrer i Cancho R., Capocci A., Caldarelli G.: Spectral methods cluster words of the same class in a syntactic dependency network Int. J. Bifurcation Chaos. 17 2453–2463 (2007)
7. Garg, M., Kumar, M.: Identifying influential segments from word co-occurrence networks using AHP. Cognitive Systems Research 47, 23–41 (2018)
8. Otter N., Porter M. A., Tillmann U., Grindrod P. and Harrington H. A.: A roadmap for the computation of persistent homology. EPJ Data Science 6(17): https://doi.org/10.1140/epjds/s13688-017-0109-5 (2017)
9. Petri G, Scolamiero M, Donato I, Vaccarino F Topological Strata of Weighted Complex Networks. PLoS ONE 8(6): e66506. doi:10.1371/ journal.pone.0066506 (2013)
10. Radhakrishnan S., Erbis S., Isaacs J. A., Kamarthi S.: Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. PLoS ONE 12(3): e0172778. https://doi.org/10.1371/journal. pone.0172778

In-block Nested Structural Patterns in Ecological and Social Networks

Claudio J. Tessone[1], Albert Solé-Ribalta[2] Manuel S. Mariani[3], and Javier Borge-Holthoefer[2]

[1] URPP Social Networks, Universität Zürich, Switzerland
claudio.tessone@uzh.ch,
[2] IN3, Universitat Oberta de Catalunya, Spain
[3] Physics Department, Université de Fribourg, Switzerland

1 Introduction

The concept of *nestedness* was coined in ecology to characterise the spatial distribution of biotas in isolated, yet spatially-related landscapes, and later found to describe large families of inter-species relations. In structural terms, a perfectly nested pattern is such that the set of interactions of any given node is a nested subset of the connections of larger degree nodes; see Fig. 1(left). Nestedness has imposed itself as a landmark feature in mutualistic classes in the most variegated disciplines.

On the other side, the identification of *modular* patterns in networks stands as one of the hallmarks in the area with prominent precedents in social network analysis. Besides social systems, networks with significant community structure, see Fig. 1 (middle), appear in multiple contexts, like biology or cognitive science. It implies the existence of subgroups of nodes, strongly connected within but loosely connected to nodes outside. The identification and analysis of community structure constitutes itself a sub-area of network science. It poses challenges with respect to detection algorithms, empirical problems, applications and conclusions derived.

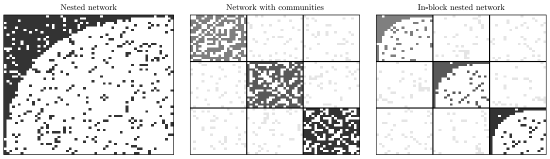


Fig. 1. From left to right. Example of network with nested structure. Rows and columns have been ordered by degree. Example of a network with community structure. High internal connectivity between nodes of the same community and low connectivity between nodes on different communities. Example of a network with in-block nested structure. Nodes within communities exhibit nested structure.Vertical and horizontal lines are a visual guide to show the existing modules.

Nestedness and modularity have been often treated as incompatible architectures, since they are thought to emerge from conflicting (cooperative, competitive) dynamics.

Thus, most studies have only focused exclusively on either of them. In this work, we introduce a compact methodological framework that jointly considers both patterns block structure and In-Block Nestedness (INS). Our findings indicate that these previously-overlooked structures are in fact common in ecological and social systems. This opens a new direction for the structural analysis of ecological and social systems and, at the same time, calls for new models to explain how such structures emerge.

2 Results

Without loss of generality consider a bipartite network describing a relationship between two sets $G = \{s, t, \dots\}$ and $\Gamma = \{\sigma, \tau, \dots\}$ with cardinalities N_r and N_c respectively. The bipartite network can be represented as a binary adjacency matrix $\mathbf{A}$ whose elements are $A_{s,\tau} = 1$ if a relationship between elements $s \leq N_r$ and $\tau \leq N_c$ exists, zero otherwise. We consider that both sets of nodes are partitioned into C disjoint subsets, termed *blocks*. This implies that for each node i, it is possible to define a membership variable α_i. Based on this, the total size of block ℓ can be obtained as $C(\ell) = \sum_{s=1}^{N_r} \delta(\alpha_s, \ell) + \sum_{\sigma=1}^{N_c} \delta(\alpha_\sigma, \ell)$, where δ is the Kronecker Delta. In addition, $C_s = \sum_t \delta(\alpha_s, \alpha_t)$ and $C_\sigma = \sum_\tau \delta(\alpha_\sigma, \alpha_\tau)$ give, respectively, the number of nodes in the block nodes s and σ belong to.

We introduce the in-block nestedness fitness $\mathscr{I}$, which quantifies to which extent a network exhibits INS,

$$\mathscr{I} = \frac{2}{N_r + N_c} \left\{ \sum_{s,t}^{N_r} \left[\frac{O_{s,t} - \langle O_{s,t} \rangle}{k_t\,(C_s - 1)} \Theta(k_s - k_t) \delta\,(\alpha_s, \alpha_t) \right] + \sum_{\sigma,\tau}^{N_c} \left[\frac{O_{\sigma,\tau} - \langle O_{\sigma,\tau} \rangle}{k_\tau\,(C_\sigma - 1)} \Theta(k_\sigma - k_\tau) \delta\,(\alpha_\sigma, \alpha_\tau) \right] \right\}. \tag{1}$$

where k_i corresponds to the degree of the element i (regardless on whether it belongs to G or Γ); $\Theta(\cdot)$ is the Heaviside step function (such that the only contributing terms are those in which the outer index has larger degree than the inner); $O_{\cdot,\cdot}$ measures the degree of overlap between row and column pairs as: $O_{s,t} = \sum_{\upsilon=1}^{N_r} A_{s\upsilon} A_{t\upsilon}$, and $O_{\sigma,\tau} = \sum_{u=1}^{N_c} A_{u\sigma} A_{u\tau}$. Null models $\langle O_{s,t} \rangle$ and $\langle O_{\sigma,\tau} \rangle$ gauge the expected overlap between a pair of nodes belonging to a class and aims at compensating the nestedness that can be explained solely by the nodes' degrees. The special case of a single block (i.e. global nestedness) is denoted $\mathscr{N}$. We have performed extensive analyses on synthetic benchmark graphs that show that optimisation of Eq. 1 leads to the correct detection of INS structures, whereas modularity optimisation or community detection algorithms do not.

In Fig. 2A, two colour-coded scatter plots are shown for uni- (left) and bipartite (right) networks. Strikingly, modularity Q and in-block nestedness $\mathscr{I}$ are not strongly correlated in real datasets. Networks that exhibit small or intermediate values of modularity (compatible with those of a random network) may show high $\mathscr{I}$, regardless of the $\mathscr{N}$ score. Also, large values of Q – which unsurprisingly display nestedness ≈ 0 – indeed exhibit both large and small $\mathscr{I}$ scores as well. Beyond the uncorrelated behaviour between the three descriptors, what surfaces here is the fact that when analysing data

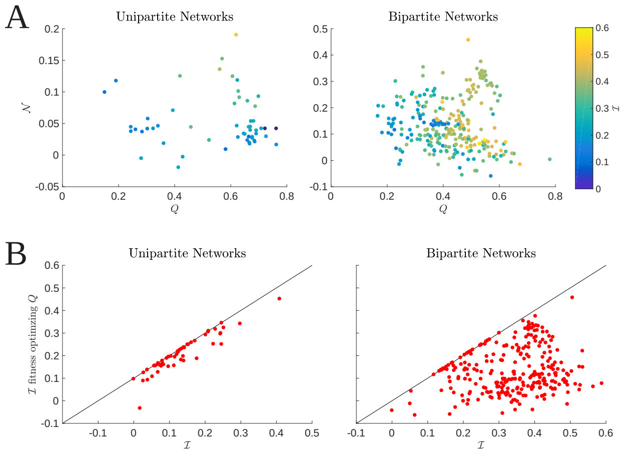


Fig. 2. (A) Scatter plot confronting modularity and nestedness with the $\mathscr{I}$ measure (color coded). Each point represents either an ecological, urban or social network (left panel: unipartite networks; right panel: bipartite networks). Note that bipartite networks have been analysed under the formulation of Barber's modularity. (B) Comparison of in-block nestedness value obtained optimising modularity or directly $\mathscr{I}$, as we propose.

we may be overlooking a relevant pattern – IBN structure – just because two partial views of it ($\mathscr{N}$ and Q taken independently) appear to be non-significant.

Further, in Fig. 2B we show the value of $\mathscr{I}$ for partitions obtained by maximising Q confronted to the maximisation of $\mathscr{I}$ itself. This plot evidences that modularity optimisation may sometimes render partitions which do have some in-block organisation (near the diagonal), but most often it is blind to it. This result highlights that using an approach where modularity is maximised, to successively evaluate nestedness within the blocks identified, is not able to unveil the IBN structure in most real-world networks.

Summary. In the past years the concept of nestedness has clearly overflowed the classical ecological framework: beyond mutualistic networks, we have now evidences that nested patterns appear in diverse settings, from anthropology and sociology to economy and urban science. Parallel to the discovery of new instances of nested organisations, scholars have debated around the co-existence of two apparently incompatible macroscale architectures: nestedness and modularity. In this regard, the discussion is far from a solution mainly for two reasons. First, nestedness and modularity appear to be the result of two contradictory dynamics, cooperative and competitive ones. Second, existing methods to evaluate the presence of nestedness and modularity are flawed when it comes to the evaluation of concurrently nested and modular structures. In this work, we define the concept and formulation of In-block Nestedness as a structural measure that assesses to what extent a network is composed of modules whose internal connectivity exhibit a nested structure. Further, we have developed a method that allows us to identify such structures successfully, both in synthetic and real networks.

Core-periphery or decentralized? Topological shifts of specialized information on Twitter

Marco Bastos[1], Carlo Piccardi[2], Michael Levy[3], Neil McRoberts[4], and Mark Lubell[3]

[1] Department of Sociology, City, University of London, UK
[2] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Italy
[3] Department of Environmental Science and Policy, University of California at Davis, USA
[4] Plant Pathology Department, University of California at Davis, USA

1 Introduction

Social media literature has long debated whether Twitter is an information diffusion system, characterized by a skewed distribution of links and low rate of reciprocal ties [1, 6], or a social network, structured around social relations, with a higher incidence of reciprocal ties and a distribution of outgoing links similar to that of incoming links [5]. The debate hinges on the overall network structure observed on Twitter and is relevant to organizations and users seeking to optimize the reach of their message in the social network.

Here we investigate shifts in Twitter network topology resulting from the type of information being shared. We identify communities matching areas of agricultural expertise and measure the core-periphery centralization of network formations resulting from users sharing generic versus specialized information. We found that centralization increases when specialized information is shared and that the network adopts decentralized formations as conversations become more generic. The results are consistent with classical diffusion models positing that specialized information comes with greater centralization, but they also show that users favor decentralized formations, which can foster community cohesion, when spreading specialized information is secondary.

2 Data and Methods

We start with 153 Twitter users identified by the University of California Division of Agricultural and Natural Resources (UCANR) as important sources of information on the topics of agriculture and environment. This initial set of users forms the seed nodes from which we snowball data collection to the larger group of users following or being followed. Our resulting dataset restricts to messages posted in 2014. We rely on the sampled data to graph a network of @-mentions and retweets connecting users. In both cases, we draw an @-mention or retweet edge connecting two accounts that have posted at least one message with specialized information at some point in 2014. These sampling processes renders a network of 4.4M edges and 32K nodes.

We rely on the approach introduced by [4] to estimate the core-periphery structure [3] of a network. By elaborating the dynamics of a random walker, a curve (the core-periphery profile) and a numerical indicator (the core-periphery score C) are derived.

T ganized in a centralized f ally, a coreness value is a left).

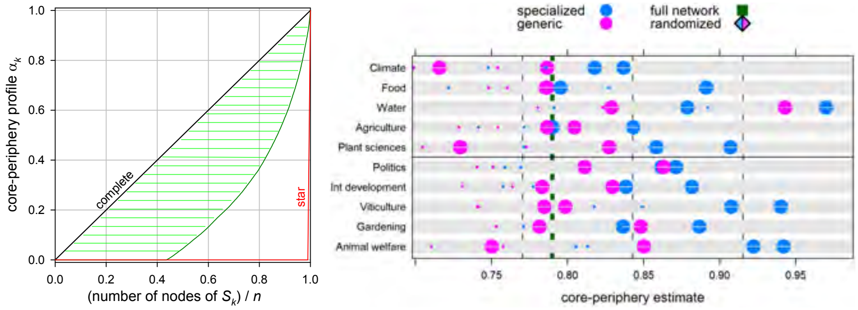


Fig. 1. Left: The core-periphery score C is the area between the core-periphery profile of a given network and that of the complete (all-to-all) network (shaded area in the figure). The value is normalized to $0 \leq C \leq 1$, so that $C = 0$ for the complete network and $C = 1$ for the star network. **Right:** Core-periphery scores C of specialized (blue) and generic (pink) hashtag-based subnetworks (large dots).

3 Results

We start by identifying communities and categorizing them into specialized topical subnetworks. Consistent with previous results [2], we find that the 10 detected communities tweet dominant hashtags that distinguish substantive thematic communities: climate change, food policy, water management, agriculture, plant sciences, politics, international development, viticulture, gardening, and animal welfare. We subsequently run multiple core-periphery analyses to test the hypothesis that higher estimates of core-periphery structure are to be observed when, within each of the 10 communities, the discussion shifts from specialized, agriculture-driven versus generic, non-agriculture-related information (we rely on an unsupervised text classifiers to identify messages as agriculture relevant). Based on such classification, we generate two subgraphs (aggie and non-aggie) of comparable size for each community and calculate the core-periphery score C. We find that the score of aggie subnetworks is larger in 10 cases out of 10, a result which is strongly indicative of the overall impact of specialized information to the network structure formed by the interaction of Twitter users.

Then, for each of the 10 communities, we select four hashtags (from the ten most frequently used) that are particular to that community, two of which are judged to be very specialized and two very generic. Therefore, two of the hashtag-based subgraphs refer to specialized conversations and two refer to generic topics of conversation within that community. We repeat the procedure for each community, hence rendering 10 × 4 subgraphs – 10 × 2 of specialized conversations and 10 × 2 of generic interactions. This shares a resemblance with the previous reported experiment, but it is specifically designed to test the hypothesis that the network structure of specialized, hashtag-based subnetworks is more centralized and that the network is increasingly structured around a core and a periphery as the topic of conversation becomes more specialized.

The results of this last experiment confirm that increasing specialization of topics is associated with star-shaped network formation, with the specialized hashtag-based subgraphs exhibiting significantly higher centralization than their generic counterparts. Figure 1 (right) unpacks these results: subgraphs of specialized conversation present consistently higher core-periphery scores C compared with the subgraphs of generic conversations. Indeed, the average core-periphery score for generic subgraphs is larger than that of specialized subgraphs (the statistical hypothesis that the two means are equal can be rejected at less than 1% significance level). Moreover, the pairwise comparison, community by community, of the mean of the two specialized scores against the mean of the two generic scores yields a larger value for the former in all 10 cases.

In conclusion, the results reported herein indicate that the communities providing specialized agricultural information become significantly more star-shaped as users tweet, retweet, or comment on messages relaying specialized information. The shift from more horizontal, decentralized topologies in which users interact and discuss generic topics towards a hierarchical, star-like structures in which information cascades from a few accounts to a large crowd of peripheral users is consistent with classical diffusion models, which posit that decentralized diffusion system are more likely to emerge when the innovations being diffused does not require high levels of technological expertise. As such, Twitter networks become more centralized and structured as an efficient information broadcast system when users change their conversation from generic to specialized topics.

References

1. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: Quantifying influence on twitter. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. pp. 65–74. WSDM '11 (2011)
2. Bastos, M.T., Puschmann, C., Travitzki, R.: Tweeting across hashtags: Overlapping users and the importance of language, topics, and politics. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. pp. 164–168. HT '13 (2013)
3. Borgatti, S., Everett, M.: Models of core/periphery structures. Social Networks 21(4), 375–395 (1999)
4. Della Rossa, F., Dercole, F., Piccardi, C.: Profiling core-periphery network structure by random walkers. Scientific Reports 3, 1467 (2013)
5. Newman, M., Park, J.: Why social networks are different from other types of networks. Physical Review E 68(3, Part 2), 036122 (2003)
6. Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who says what to whom on twitter. In: Proceedings of the 20th International Conference on World Wide Web. pp. 705–714. WWW '11 (2011)

Text networks: foundations and structural analysis

Davide Vega[1] and Matteo Magnani[1]

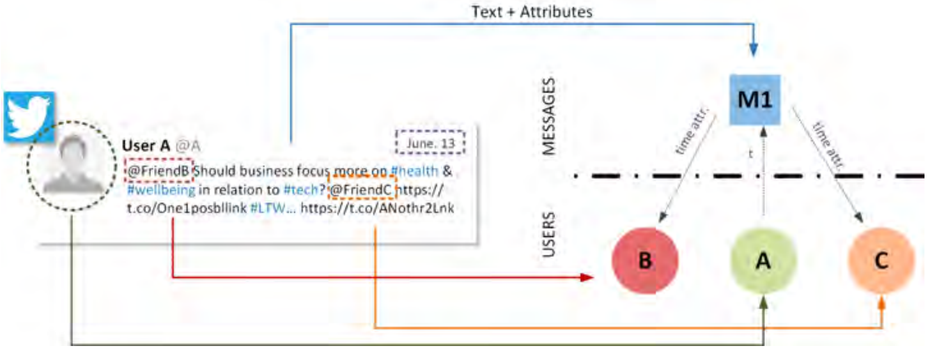
Department of Information Technology, Uppsala University, Uppsala, Sweden
davide.vega@it.uu.se
matteo.magnani@it.uu.se

1 Introduction

A large amount of human generated information is available online in the form of text exchanged between individuals or groups. Examples include social network sites, online forums and emails. The public accessibility of several of these sources allows us to observe our society at various scales, from conversations among small groups of individuals to the effects of misinformation on large communities.

To cope with the complexity of online information, researchers have typically focused on either the topology of the network, as commonly done in Network Science, or the text exchanged among individuals, using methods from Computational Linguistics. In both cases time has also been taken into consideration, as in Temporal Networks or Temporal Information Retrieval.

In this work, **we introduce an attributed multilayer model for temporal text networks**, enabling the application of a wide range of existing methods to this context. Our model can represent all the information contained in the aforementioned data sources, including different types of text interactions, such as direct messages exchanged between individuals, multicast information targeting specific communities or broadcast news.



@FriendB Should business focus more on #health & #wellbeing in relation to #tech? @FriendC https://t.co/One1posbllink #LTW... https://t.co/ANothr2Lnk

Fig. 1. Codification example of online information from Twitter.

We also introduce two main approaches to analyze text networks, that we call *discrete* and *continuous*. In the discrete model text messages can be classified into a number of (possibly overlapping) groups. In the continuous approach, a comparison function generates a continuous score indicating the similarity of the text messages (e.g., based on their content).

Finally, we introduce various new types of projections based on the discrete and continuous models. A projection is a generic multilayer network operator that creates edges in one layer based on the information present in another layer. We show that starting from our unified multilayer model of text networks we can project its information into several derived types of networks, such as communication networks, user-annotated networks, topic-based multiplex networks and information propagation networks for which existing analysis methods exist.

2 Results

To exemplify the potential of our model, we will present the results of the analysis of a Twitter dataset containing all the tweets with hashtag *#LTW*, which was the main hashtag used during the London Tech Week — a major technological conference held in London from June 12 to June 16, 2017. By coding each tweet as in Fig. 1, we obtain a multilayer network with 934 users, their follower/followed relations and the 4898 messages exchanged between them (See Fig. 2 *left*).

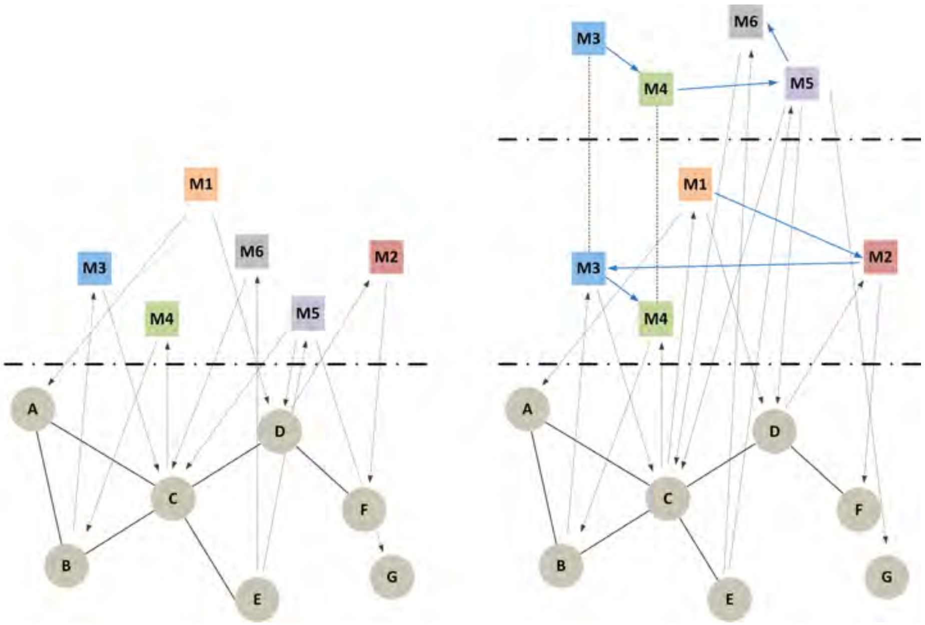


Fig. 2. Attributed multilayer model for temporal text networks (*left*) and one possible projection of the text layer based on topic analysis (*right*).

We can now apply a text discretization by identifying the topics of the tweets, so that each topic is projected into a separate layer (See Fig. 2 *right*). This creates a multiplex network where interactions on different topics are coded into different layers. In our working example we have identified the topics by weighting and clustering the hashtags, obtaining as a result: *education*, *womenintech*, *healthcare*, *hackathon*, *smartcities*, *technology*, *startups*, *ai*, *financial*, *iot*, *menabling17*, *drones*, *industry*, *telecom*, *cloud*. So, our example tweet will be represented as two interconnected nodes present in the *healthcare* and *technology* layers.

At this point we can apply standard multilayer methods [4] [3], obtaining as a result information about the three components: the network structure, the topics and their time evolution. For example, using the abacus [1] community detection algorithm for multilayer networks we are able to identify not only communities of users strongly connected or sets of tweets about similar topics, but also topic-specific discussion groups.

For space reasons, here we have described a specific data analysis workflow, starting from our general data model, applying a text discretization function, a number of projections and finally a multilayer clustering method. Using different types of projections, or working directly with the original model, several types of text and structural analysis can be unified under the same framework. Examples include conversation retrieval [5], where a message ranking method is defined as a function of text, relationships between messages and between senders/recipients, and temporal information, and text message clustering [2] — as opposed to our working example where we cluster people using the information exchanged between them.

References

1. Berlingerio, M., Pinelli, F., Calabrese, F.: ABACUS: frequent pAttern mining-BAsed Community discovery in mUltidimensional networkS. Data Mining and Knowledge Discovery 27(3), 294–320 (2013)
2. Combe, D., Largeron, C., od Egyed-Zsigmond, E., Géry, M.: Combining relations and text in scientific network clustering. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 1280–1285. Istanbul, Turquie (2012)
3. Dickison, M.E., Magnani, M., Rossi, L.: Multilayer Social Networks. Cambridge University Press (2016)
4. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer Networks. Journal of Complex Networks 2(3), 203–271 (sep 2014)
5. Magnani, M., Montesi, D., Rossi, L.: Conversation retrieval for microblogging sites. Information Retrieval 15(3-4), 354–372 (feb 2012)

tikz-network: a LaTeX library for vizualizing complex networks

Jürgen Hackl[1]

Swiss Federal Institute of Technology (ETH), Zurich, Switzerland,
hackl@ibi.baug.ethz.ch

Abstract. `tikz-network` is an open source software project for visualizing graphs and networks in LaTeX. It aims to provide a simple and easy tool to create, visualize and modify complex networks. The packaged is based on the PGF/Ti*k*Z languages for producing vector graphics from a geometric/algebraic description. Particular focus is made on the software usability and interoperability with other tools. Simple networks can be directly created within LaTeX, while more complex networks can be imported from external sources (e.g. igraph, networkx, QGIS, ...). Additionally, `tikz-network` supports visualization of multilayer networks in two and three dimensions. The software is available at:
https://github.com/hackl/tikz-network.

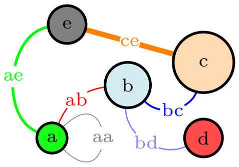
1 Introduction

In recent years, complex network theory becomes more and more popular within the scientific community. Besides a solid mathematical base on which these theories are built on, a visual representation of the networks allow communicating complex relationships to a broad audience.

Nowadays, a variety of great visualization tools are available, which helps to structure, filter, manipulate and of course to visualize the networks. However, they come with some limitations, including the need for specific software tools, difficulties to embed the outputs properly in a LaTeX file (e.g. font type, font size, additional equations and math symbols needed, ...) and challenges in the post-processing of the graphs, without rerunning the software tools again.

In order to overcome this issues, the package `tikz-network` was created. Since LaTeX is a standard for scientific publications and widely used, there is a high chance that users are already familiar with the syntax and the structure of this language. Beside LaTeX, no other software tool is needed. The commands of `tikz-network` are kept simple but allow a high control over the produced output. Post-processing of the network (e.g. adding drawings, images, texts, equations,...) can be done easily, due to the compatibility with PGF/Ti*k*Z [1]. Also, the embedding of the network visualization into the LaTeX-environment enables the use of the fonts, font sizes, mathematical symbols, hyperlinks, references,..., as used in the document. Additional features are the three-dimensional visualization of (multilayer) networks, and the compatible with other layout and visualization tools (e.g. igraph, netwrokx, QGIS,...)

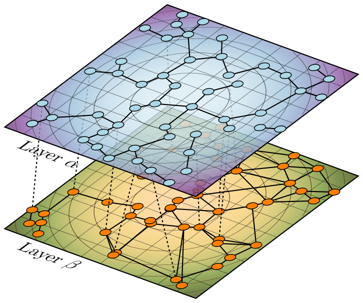
2 Examples



```
\begin{tikzpicture}
  \Vertex[size=.4,color=green,opacity=.9,label=a]{A}
  \Vertex[x=1,y=.7,opacity=.5,label=b]{B}
  \Vertex[x=2,y=1,size=.8,color=orange,opacity=.3,label=c]{C}
  \Vertex[x=2,size=.5,color=red,opacity=.7,label=d]{D}
  \Vertex[x=.2,y=1.5,size=.5,color=gray,label=e]{E}

  \Edge[label=ab,lw=.5,color=red,bend=30](A)(B)
  \Edge[label=bc,lw=.7,color=blue,bend=-60](B)(C)
  \Edge[label=bd,lw=.5,color=blue,opacity=.5,bend=-60](B)(D)
  \Edge[label=ae,lw= 1,color=green,bend=75](A)(E)
  \Edge[label=ce,lw= 2,color=orange](C)(E)
  \Edge[label=aa,lw=.3,color=black,opacity=.5,bend= 75](A)(A)
\end{tikzpicture}
```

Fig. 1. Example: simple network without external input



```
\SetCoordinates[xAngle = -20]
\begin{tikzpicture}[multilayer=3d]
  \SetVertexStyle[MinSize = 4.5mm]
  \SetLayerDistance{-5}
  \SetPlainWidth{10}
  \SetPlainHeight{10}

  % Layer beta
  \Plain[layer=2,color=orange,opacity=.6,image=field_b,
        ImageAndFill,grid=1cm,InBG]
  \Text[x=1.2,y=-.1,layer=2,anchor=north west,style={scale
        =2.5}]{Layer $\beta$}
  \Vertices[color=orange]{V.csv}
  \Edges[color=black,layer={2,2}]{E.csv}
  \EdgesNotInBG
  \Edges[color=black,layer={1,2},style={dashed}]{E.csv}

  % Layer alpha
  \Plain[opacity=.6,image=field_a,ImageAndFill,grid=1cm]
  \Text[x=1.2,y=-.1,layer=1,anchor=north west,style={scale
        =2.5}]{Layer $\alpha$}
  \Edges[color=black,layer={1,1}]{E.csv}
  \Vertices[layer=1]{V.csv}
\end{tikzpicture}
```

Fig. 2. Example: mutlilayer network with external vertex and edge data

3 Future work

Though the core of the package already exists, further work is required for the development of new features and tools. A special focus is made on interfaces between other network visualization tools.

Since `tikz-network` is an open source software project, any community feedback or suggestion are welcome, in order to improve the library and adapt to the needs of the users.

References

1. Tantau, T.: The tikz and pgf packages. Manual Version 3.0.1a, Universität zu Lübeck, Lübeck, Germany (8 2015), http://sourceforge.net/projects/pgf

Dynamical efficiency in congested road networks

Leonardo Bellocchi and Nikolas Geroliminis

Ecole Polytéchnique Fédérale de Lausanne (EPFL), Laboratory of Urban Transportation Systems (LUTS), Lausanne, Switzerland
leonardo.bellocchi@epfl.ch

Keywords: Efficiency, complex networks, shortest time path, urban traffic, congestion propagation.

1 Introduction

The main topological properties of a transportation network can be characterized using different criteria such as structure, the degree distribution of nodes and connectivity. In literature ([2]), *Efficiency* is a property of a network that measures the 'straightness' for a walker in a graph to reach nodes in the network. Efficiency, in its static form, has been widely utilized in several spatial networks like communication, urban and physical networks. In this study, we aim to apply a similar definition for transportation network and compute the changes over time of this corresponding connectivity measure, that we called dynamical efficiency, considering average link speed in a city during a congestion period. We show how these new measures of traffic performance at the network level provide a better understanding of congestion evolution compared to state of the art measures and can be utilized to evaluate various management techniques. It describes not the simple link congestion level but how the link is good for the mobility in relation with the network also in heavy traffic condition. Results analysis (i) of a large number of taxis from a megacity in China are presented.

2 Efficiency and betweenness: from static to time depending definition

Let $\mathscr{G}(N,V)$ be a graph of N nodes and V links that represents a spatial network. According to the literature ([2]) we could define *efficiency* for $\mathscr{G}$ like average ratio between the euclidean distance and the shortest path between each pair of nodes in N.

The main idea of this paper is to extend this analysis to a dynamical case where the length of links can change or, like in the transportation network, the average speed of links may vary in time because of some congestion. This is not common in the literature of networks, where the link speeds are fixed (see for example [2]) or, like in [6], network disruptions are considered but not their degradation.

In order to take into account this effect we define ([1])

$$E(t) = \frac{1}{N(N-1)} \sum_{i,j \in N} \frac{\tau_{ij}^{FF}}{\tau_{ij}^{STP}(t)} \qquad , \qquad E(i,t) = \frac{1}{N-1} \sum_{j \neq i \in N} \frac{\tau_{ij}^{FF}}{\tau_{ij}^{STP}(t)}$$

resp. *global dynamical efficiency* and *local dynamical efficiency* at time t for link i, where τ_{ij}^{FF} is the shortest time path (STP) between node i and j under free flow scenario and $\tau_{ij}^{STP}(t)$ is the STP calculated with the experienced speed at time t, i.e. using the average link speeds from available data. In the full paper, we will also take into consideration the changes in the distribution of betweenness centrality and so, the use of a road network under traffic congestion.

3 Results

We have tested our algorithm with some data of the traffic condition of Shenzhen (China) during the morning peak hour (6am - 8am). Our available dataset consists on link speed estimations based on GPS signals (with a frequency of 30s) of about 20k taxis active in the monitored period (for more detail see [4]). For each link and every 5 minutes we have the average speed of all taxis passed through it in the last 5 minutes. Thanks to this speed estimation we compute the travel time for each link l in the network, that is $travel\_time(l,t) = link\_length(l)/speed(l,t)$ from 6am to 8am (in total 25 speed records for each of the 580 links). We, therefore, compute the *local dynamical efficiency* $E(i,t)$ for each node i and every 5 minutes window $t = 6am, 6:05am, \ldots, 8am$. Then, in Figure 1 we assigned to each link the average value of its two extreme nodes.

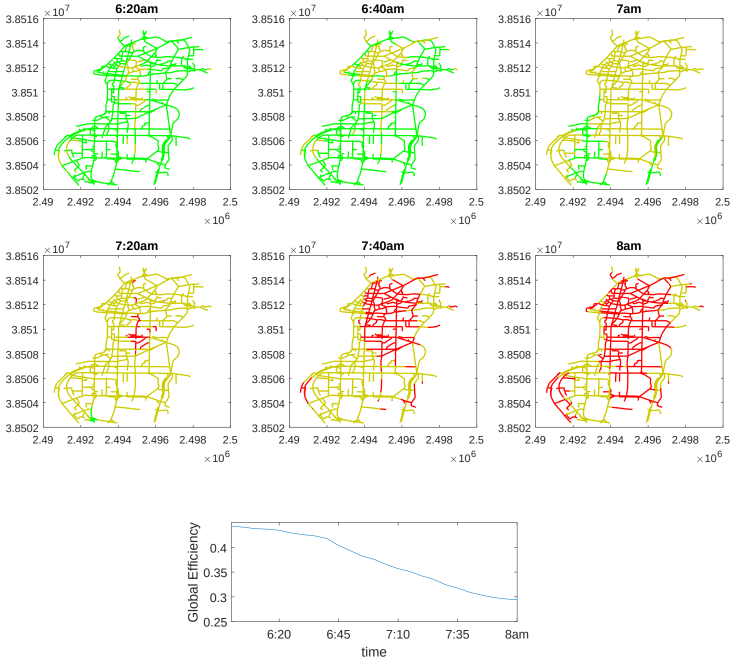


Fig. 1. Contour plot of local D-efficiency for Shenzhen downtown during the morning peak hour every 20min form 6:20am to 8am from top-left to bottom-right. The link are colored based on their D-efficiency in tree equal categories: green (high D-efficient $E > 0.4$), yellow (average D-efficient), red (low D-efficient $0 < E < 0.33$). In the bottom panel the evolution of the average D-efficiency during the peak hour.

This type of analysis results seems to be suitable and befitting for typical transportation data, like GPS points, loop detectors, etc.. In fact, we show that even if data are limited and/or noisy the global picture of the networks efficiency is still consistent. This is because for each node i are considered all shortest time path from i to every other node j in the network, and so the value of its *local dynamical efficiency* is averaged and does not depend directly just on its speed value. The other important feature that with this measure we highlight is the clear appearance of the connected congested components and their propagation through the city (see Figure 1). This last fact allows, among others, traffic engineers to intervene with perimeter controls techniques in order to mitigate the traffic (see for example [5]). Due to the characteristic of this measurement, the city results divided into clusters of congestion with no need of costly clustering algorithms.

4 Heuristics

All our results derive in an immediate way from a unique simple algorithm, that is the computation of STP for each pair of nodes and for each time slot (commonly 5-15 minutes each) we are interested in. This imply that in large networks, or for a long time period, the computational cost increases exponentially.

In order to avoid this cost, we consider how to exploit the specific dynamics of the urban traffic in, at least, two solutions that can still give a good representation of traffic conditions with less computational effort. As first approach, we can compute only the STP between all pairs of nodes distanced less than r meters. It means to compute the STP from each node $i = 1,\dots,n$ to every node $j \in N_i(r) = \{j \in N : d(i,j) \leq r\}$. In order to guarantee that $N_i(r) \neq \emptyset$ for each $i \in N$ we have to take care that $r > \max_i \min_j d(i,j)$.

Moreover, the results shown in the full paper confirm that the dynamical efficiency depends mainly on the local traffic condition. This max-radius approximation is supported also by some recent papers (for example [3]) that show that the length trip probability distribution in a city follows the law ($P(r) = \frac{1}{r^\gamma}$ with $\gamma \approx 3$) and so it makes sense to stop the computation of the STP for a radius r big enough to consider the most part of them.

A second approach is based on the knowledge of the origin-destination matrix demand. The main idea is to take into account this information and give a proportional weight w_{ij} at their estimated demand in the efficiency formula for each pair of nodes (i,j), that is $E_w(i,t) = \frac{1}{\sum_{j \neq i \in N} w_{ij}} \sum_{j \neq i \in N} w_{ij} \frac{\tau_{ij}^{FF}}{\tau_{ij}^{SP}(t)}$. The advantages of this approach is that very often the demand between many nodes is zero or negligible. The decision to take into account only the OD pair reduces the computational cost and also provides a measure of efficiency of urban system not only in term of network but in term of users.

References

1. Bellocchi, L., Geroliminis, N.: Dynamical efficiency in congested road network. 16th Swiss Transport Research Conference (2016)
2. Crucitti, P., Latora, V., Porta, S.: Centrality measures in spatial networks of urban streets. Phys. Rev. E 73, 036125 (2006)
3. Gallotti, R., Bazzani, A., Rambaldi, S., Barthelemy, M.: A stochastic model of randomly accelerated walkers for human mobility. Nature Communications 7, 12600 (2016)
4. Ji, Y., Luo, J., Geroliminis, N.: Empirical observations of congestion propagation and dynamic partitioning with probe data for large-scale systems. Transportation Research Record (2422), 1–10 (2014)
5. Kouvelas, A., Saeedmanesh, M., Geroliminis, N.: Enhancing model-based feedback perimeter control with data-driven online adaptive optimization. Transportation Research Part B: Methodological 96, 26–45 (2017)
6. Scellato, S., Leontiadis, I., Mascolo, C., Basu, P., Zafer, M.: Evaluating temporal robustness of mobile networks. IEEE Trans. Mob. Comp. 12(1), 105–117 (2013)

Part X

Community Structure

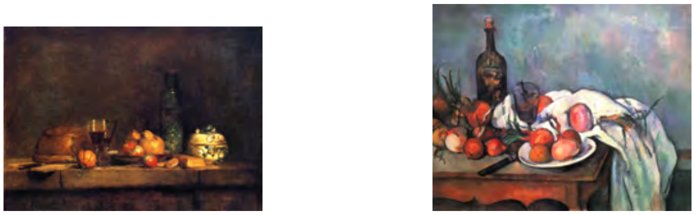
Community Detection with Metadata in a Network of Artistic Influence

Michael Kitromilidis and Tim S. Evans

Centre for Complexity Science and Theoretical Physics Group
Imperial College London, SW7 2AZ, UK.
m.kitromilidis14@imperial.ac.uk

1 Introduction

In this paper we ask two questions: who is the most important painter and who was the most influential. For example, many people will recognise the impressionist Paul Cézanne as a famous master in the history of art. Sources of his inspiration can be traced two hundred years back, in the Baroque painter Jean-Baptiste-Siméon Chardin (less well known, but clearly highly influential), who belonged in a completely different artistic movement and period.



(a) Chardin (1728) (b) Cézanne (1898)

Table 1. Example of how Chardin's still life works were influential for impressionists, such as Cézanne. Note the angled knife on the left of the paintings, used by Chardin as a trick to give a sense of depth in his paintings, also observed in Cézanne's work.

In our work we wish to approach artistic influence from a networks perspective. We begin by forming a new dataset of Western art painters, using data from the Web Gallery of Art (www.wga.hu) and Wikipedia. Similar studies of networks have argued that community structure in person networks is strongly associated with professional similarities, such as the work by Gleiser and Danon on jazz musicians [3] or the work by Goldfarb et al. which also performs Wikipedia analysis [4].

Our first question on importance of painters can be studied using standard centrality measures. However, our investigation of influence led us to produce some new network tools. First we had to adapt network community detection methods to produce better results by exploiting metadata as well as network topology. Secondly, we compared community aware centrality measures with traditional centrality measures in order to identify influential nodes, as we argue that to identify truly influential nodes in a network we have to look beyond the cluster where they belong.

2 Methods

Community detection in social networks has typically been performed using external metadata to a certain extent. While in early studies metadata were used as a ground truth in order to assess the quality of a community partition, in more recent works metadata are used together with the network structure to identify a partition of a network into communities [5–7].

In social networks metadata refer to attributes and characteristics that the nodes have, including their personal, professional, demographic and social identities. In our context of Western art painters, metadata are used to detect communities corresponding to artistic genres and the countries where they were active. A standard Louvain modularity [1] split into communities reveals clusters already centred largely (but not exclusively) around artistic genres, see Figure 1.

Our aim is to take metadata into account to assess this partition. More generally, we consider that every node in a network has an associated vector of metadata, $X_i = \left(x_1^{(i)}, \ldots, x_n^{(i)}\right)$. Given a partition of a network $\mathscr{C}$ into communities $\{c_\alpha\}$, we define two measures. The first, *cluster homogeneity*

$$h_{\mathscr{C}}(c) = \frac{1}{M} \sum_{i,j \in c} S(X_i, X_j) \tag{1}$$

is a measure of how similar the metadata of nodes in a particular cluster c are, where M is cluster size and $S(x,y)$ an appropriate similarity function. The second measure, *configuration entropy*

$$e_{\mathscr{C}}(\xi) = - \sum_{c_\alpha \in \mathscr{C}} p_\alpha(\xi) \log p_\alpha(\xi) \tag{2}$$

measures how fragmented a specific configuration vector $\xi \in \mathbb{R}^n$ of metadata is, where p_α is the probability of finding that configuration in cluster α.

We can then use these two measures to assess whether a community partition of a network is underdetecting or overdetecting communities; in the former case we look at the structure within clusters, in the latter we look at different groupings of clusters.

In the context of our painters metadata correspond to two categories, as extracted from the WGA, the artists' movements and their activity locations, and we consider a finer partition of the network as well as the standard Louvain partition.

3 Results

Among our results we discuss how a standard and finer partition into communities help us discover which nodes in the network have influence beyond their own communities. In the context of painters this means artists whose work had an impact in different genres, possibly in different artistic periods. To quantify this we look at standard

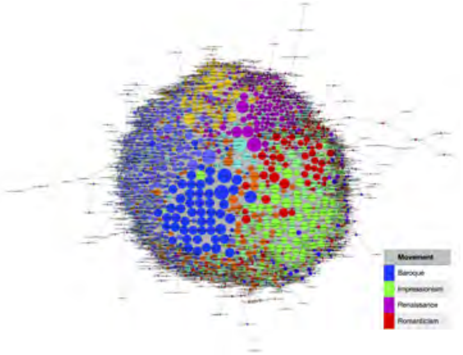


Fig. 1. Partition of the painter network into Louvain communities and correspondence of clusters with artistic movements.

community-based measures and propose new ones, variations on the standard centrality measures which take the community structure into account.

For example, the mixing parameter [2], $\mu_i = \frac{k_i^{\text{out}}}{k_i}$ is an intuitive measure in terms of the degree. We also propose looking at variants of other centrality measures, such as betweenness $bc_{\mathscr{C}}(i) = \sum_{k,l:\delta_{c(k),c(l)}=0} \frac{\sigma_{kl}(i)}{\sigma_{kl}}$ and closeness centrality $cc_{\mathscr{C}}(i) = \frac{1}{\sum_{j \notin c(i)} d_{ij}}$ where we focus only on paths starting and finishing in different communities.

By looking at nodes who score poorly in the traditional measures but highly in the modified community-aware ones, we are able to uncover a more subtle notion of influence in a person network.

Acknowledgement. MK would like to thank the A.G. Leventis Foundation for the financial support.

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10), P10008 (2008)
2. Fortunato, S., Hric, D.: Community detection in networks: A user guide. Physics Reports 659, 1–44 (2016)
3. Gleiser, P.M., Danon, L.: Community structure in jazz. Advances in complex systems 6(04), 565–573 (2003)
4. Goldfarb, D., Merkl, D., Schich, M.: Quantifying cultural histories via person networks in wikipedia. arXiv preprint arXiv:1506.06580 (2015)
5. Hric, D., Darst, R.K., Fortunato, S.: Community detection in networks: Structural communities versus ground truth. Physical Review E 90(6), 062805 (2014)
6. Peel, L., Larremore, D.B., Clauset, A.: The ground truth about metadata and community detection in networks. Science Advances 3(5), e1602548 (2017)
7. Yang, J., McAuley, J., Leskovec, J.: Community detection in networks with node attributes. In: Data Mining (ICDM), 2013 IEEE 13th international conference on. pp. 1151–1156. IEEE (2013)

Community characterization in a large-scale Japanese production network

Abhijit Chakraborty*, Hazem Krichene, Hiroyasu Inoue, and Yoshi Fujiwara

Graduate School of Simulation Studies, University of Hyogo, Kobe 650-0047, Japan
*abhiphyiitg@gmail.com

Production network of a country constitutes the backbone of its economy. Such a network exhibits a scale-free degree distribution, disassortative mixing and a prominent community structure [2]. Communities are the building blocks of such network and play pivotal roles in its economic activity. Previous studies on the communities in the Japanese production network, based on modularity maximization [1], have shown that the communities are characterized by firms location and industrial sector [3, 4]. However, modularity maximization technique for community detection in production network considers only the topology of the network and do not capture the dynamical behaviour in the network, as the capital flows between firms. Because the links in a production network represent the flow of capital from one firm to another, it is more appropriate to use the "Infomap" algorithm [5, 6] for community detection. Infomap algorithm capture more complex structures in the network than techniques based solely on the network topology.

We present an analysis of a production network based on a large-scale nationwide inter-firm relationship data from Japan. To define the network, each firm is treated as a node, and a directed link of the form $A \rightarrow B$ indicates that A is a supplier firm for firm B. The production network exhibits a scale-free degree distribution, hierarchical clustering and disassortative mixing, consistent with previous studies. We conduct detailed investigations to characterize the communities, which are uncovered by applying Infomap algorithm [6] on largest weakly connected component of the production network, having $N = 1,234,687$ nodes and $L = 5,481,403$ directed links. The analysis of the empirical production network reveals 311 communities and 8054 inter-community links at the top community level. To determine the statistical significance of our result, we compare it with the result obtained from an identical analysis of a 'null model', which is a maximally random network with the same nodal in- and out-degrees. In stark contrast with the empirical result, the maximally random network is found to contain $62,917 \pm 120$ communities and $4,234,729 \pm 1,552$ inter-community connections. As seen from Fig. 1 (left), the community size distributions for the empirical network and the maximally random network are found to be distinctly different in nature. Whereas the empirical network has a broad distribution of community sizes with a wide range of values spanning several orders of magnitude, the maximally random network has a comparatively narrow distribution in which the largest community size is $\sim 1,000$.

We investigate the topological features within each community in the production network. The link density within a community is the ratio of the number of internal links to the maximum possible number of links. The link density ρ within a community of size s can be calculated as $\rho = e/s(s-1)$, where e is the number of links within the community. The scaled link density $\bar{\rho}$ within a community is defined as $\bar{\rho} = \rho s = e/(s-1)$.

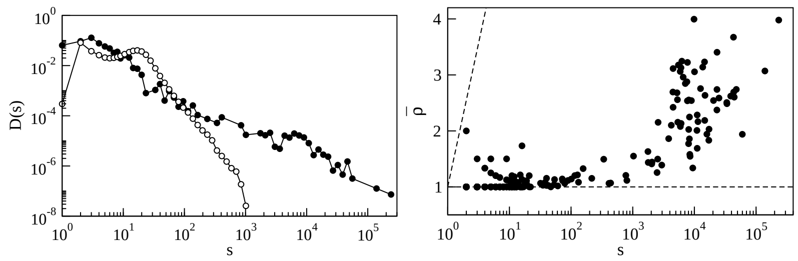


Fig. 1. (left) Distributions $D(s)$ of community sizes s for the empirical production network with directed links (filled circles) and for its randomized counterpart (open circles). Communities are detected in the largest weakly connected component of the network. Logarithmic binning is used for the horizontal axis. (right) Scatter plot of the intra-community scaled link density $\bar{\rho}$ versus the community size s. The dotted lines represent two limiting cases: $\bar{\rho} = s$ (clique) and $\bar{\rho} = 1$ (tree).

A value of $\bar{\rho} = 1$ corresponds to a community with a tree-like structure with unidirectional links, and $\bar{\rho} = s$ corresponds to a complete graph structure, i.e., a structure in which every node is connected to all other nodes in the community. Fig. 1 (right) shows a scatter plot of the scaled link density $\bar{\rho}$ versus the community size s. It is evident that the network structures within the communities are far from being complete graph structures; indeed, they are very close to an ideal tree-like structure when the community size is small ($s < 1{,}000$). However, beyond ($s > 1{,}000$), the scaled link density gradually increases as the community size increases.

Rank	Size	Over-expression of sectors	Over-expression of prefectures
1	233, 294	Manufacturing, Electronics, Water transport, Wholesale, etc.	Urban prefectures (Tokyo and its neighboring prefectures, Osaka, Aichi, Hyogo, etc.)
2	139, 380	Agriculture, Food, Fisheries, Road freight transport, Cooperative associations, N.E.C., etc.	Rural prefectures (Aomori, Miyagi, Shizuoka, etc.)
3	59, 906	Construction, Real estate, Banking, etc.	Tokyo and its neighboring prefectures, Osaka
4	47, 849	Manufacture of textiles, rubber, leather, etc.	Tokyo, Osaka, Kyoto, Aichi, etc.
5	44, 349	Medical services, Research institutes, Chemical products, etc.	Hokkaido, Tokyo, Hiroshima, etc.
6	43, 397	Retail trade (machinery and equipment), Automobile maintenance, Transport, Insurance institutions, etc.	Many (22) prefectures
7	43, 018	Multiple sectors	Hokkaido
8	38, 819	Multiple sectors	Tokyo
9	33, 654	Information services and many others	Tokyo, Kanagawa, Osaka
10	33, 563	Construction and others	Gifu, Aichi, Mie

Table 1. Brief summary of our results on the over-expression of sectors and prefectures in the ten largest communities

Our study also shows that a large fraction ($\sim 40\%$) of firms with relatively small in- or out-degrees have customers or suppliers solely from within their own communities, indicating interactions with a highly local nature. The interaction strengths between communities as measured by the inter-community link weights exhibit a highly heterogeneous distribution.

A deeper characterization of the obtained communities is derived using a rigorous statistical procedure [7] to show the over-expression of prefectures and sectors. A brief summary of the over-expression of prefectures and sectors in the ten largest communities is tabulated in table 1. We conclude that different communities are characterized by distinct features related to different sectors and prefectures.

Our results suggest that the production network exhibits many statistically significant structural patterns which might be useful to study GDP fluctuations, risk propagation and cascading failure in the network. In future, we would like to generalize the applied technique to characterize the communities at all levels.

Acknowledgement

This research was supported by MEXT as Exploratory Challenges on Post-K computer (Studies of Multi-level Spatiotemporal Simulation of Socioeconomic Phenomena). We thank Hideaki Aoyama, Hiroshi Iyetomi and Yuichi Kichikawa for useful discussions on Infomap.

References

1. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical review E 70(6), 066111 (2004)
2. Fujiwara, Y., Aoyama, H.: Large-scale structure of a nation-wide production network. The European Physical Journal B 77(4), 565–580 (2010)
3. Iino, T., Kamehama, K., Iyetomi, H., Ikeda, Y., Ohnishi, T., Takayasu, H., Takayasu, M.: Community structure in a large-scale transaction network and visualization. In: Journal of Physics: Conference Series. vol. 221, p. 012012. IOP Publishing (2010)
4. Iino, T., Iyetomi, H.: Community structure of a large-scale production network in japan. In: The Economics of Interfirm Networks, pp. 39–65. Springer (2015)
5. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105(4), 1118–1123 (2008)
6. Rosvall, M., Bergstrom, C.T.: Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PloS one 6(4), e18209 (2011)
7. Tumminello, M., Miccichè, S., Lillo, F., Varho, J., Piilo, J., Mantegna, R.N.: Community characterization of heterogeneous complex systems. Journal of Statistical Mechanics: Theory and Experiment 2011(01), P01019 (2011)

A latent geometry rationale for engineering graph-dissimilarities enhances affinity propagation community detection in real complex networks

Alessandro Muscoloni[1] and Carlo Vittorio Cannistraci[1,2,*]

[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany
[2] Brain bio-inspired computation (BBC) lab, IRCCS Centro Neurolesi "Bonino Pulejo", Messina, Italy

Affinity propagation (AP) is a state of the art algorithm for data clustering [1], however the numerous attempts to test its performance for community detection in real complex networks [2]–[5] have been attaining results very far from the state of the art methods [6] such as Infomap [7] and Louvain [8]. Yet, all these studies agreed that the crucial problem is to convert the network topology in a 'smart-enough' dissimilarity matrix that is able to properly address the message passing procedure behind affinity propagation clustering.

The former solutions were based on engineering by hand or making attempts to 'try' some topological similarity measures (converted in dissimilarity matrices) already known in network science, but these efforts were not guided by any network science concept or notion to tailor the dissimilarity measure as a geometry that favours the message passing procedure of affinity propagation. The dissimilarity measures most employed in previous studies are shortest path (SP) [3], Euclidean shortest path (ESP) [3], common neighbours (CN) [2] and Jaccard (J) [2].

Here we propose a rationale according to which the graph dissimilarity should approximate the distances on the hidden nonlinear manifold that characterizes the graph geometry [9], [10]. The fact that the network topology emerges from this hidden geometry is the reason why many networks can efficiently send messages according to a greedy routing procedure [9]. This greedy message propagation is facilitated by the hyperbolic and tree like structure of many real complex networks [9]–[12].

Our results demonstrate that the two dissimilarity topological measures inspired by our strategy [13], Repulsion-Attraction (RA) and Edge-Betweenness-Centrality (EBC), boost affinity propagation performance to levels that slightly outperform current state of the art methods for community detection. This is confirmed not only on the original real networks (Table 1), but also when their topology is perturbed by noise simulated by random deletion of links (missing topological information) (Table 2) or random addition of links (spurious topological information (Table 3).

References

[1] D. Dueck and B. J. Frey, “Clustering by Passing Messages Between Data Points,” *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[2] W. F. Guo and S. W. Zhang, “A general method of community detection by identifying community centers with affinity propagation,” *Phys. A Stat. Mech. its Appl.*, vol. 447, pp. 508–519, 2016.

[3] H. W. Liu, “Community detection by affinity propagation with various similarity measures,” *Proc. - 4th Int. Jt. Conf. Comput. Sci. Optim. CSO 2011*, pp. 182–186, 2011.

[4] Y. Shuzhong and L. Siwei, “Community detection based on adaptive kernel affinity propagation,” *Comput. Sci. Inf. Technol. 2009. ICCSIT 2009. 2nd IEEE Int. Conf.*, vol. 22, no. 2013, pp. 1–4, 2009.

[5] J. Vlasblom and S. J. Wodak, “Markov clustering versus affinity propagation for the partitioning of protein interaction graphs.,” *BMC Bioinformatics*, vol. 10, p. 99, 2009.

[6] Z. Yang, R. Algesheimer, and C. J. Tessone, “A Comparative Analysis of Community Detection Algorithms on Artificial Networks,” *Sci. Rep.*, vol. 6, p. 30750, 2016.

[7] M. Rosvall and C. T. Bergstrom, “Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems,” *PLoS One*, vol. 6, no. 4, p. e18209, 2011.

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *J. Stat. Mech Theory Exp.*, vol. 2008, no. 10, p. 10008, 2008.

[9] M. Boguñá, D. Krioukov, and K. C. Claffy, “Navigability of complex networks,” *Nat. Phys.*, vol. 5, no. 1, pp. 74–80, 2008.

[10] F. Papadopoulos, M. Kitsak, M. A. Serrano, M. Boguñá, and D. Krioukov, “Popularity versus similarity in growing networks,” *Nature*, vol. 489, no. 7417, pp. 537–540, 2012.

[11] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, “Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding,” *Bioinformatics*, vol. 29, no. 13, pp. 199–209, 2013.

[12] C. V. Cannistraci, T. Ravasi, F. M. Montevecchi, T. Ideker, and M. Alessio, “Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes,” *Bioinformatics*, vol. 26, pp. i531–i539, 2010.

[13] J. M. Thomas, A. Muscoloni, S. Ciucci, G. Bianconi, and C. V. Cannistraci, “Machine learning meets network science: dimensionality reduction for fast and efficient embedding of networks in the hyperbolic space,” *arXiv:1602.06522*, 2016.

Method	**Karate**	**Opsahl 8**	**Opsahl 9**	**Opsahl 10**	**Opsahl 11**	**Polbooks**	**Football**	**Polblogs**	**Mean**
	N=34	N=43	N=44	N=77	N=77	N=105	N=115	N=1222	
	E=78	E=193	E=348	E=518	E=1088	E=441	E=613	E=16714	
	C=0.59	C=0.61	C=0.68	C=0.66	C=0.72	C=0.49	C=0.40	C=0.36	
	γ=2.12	γ=8.20	γ=5.92	γ=5.06	γ=4.87	γ=2.62	γ=9.09	γ=2.38	
	m=2.29	m=4.49	m=7.91	m=6.73	m=14.13	m=4.20	m=5.33	m=13.68	
	N_C= 2	N_C=7	N_C=7	N_C=4	N_C=4	N_C=3	N_C=12	N_C=2	
EBC-AP	0.83	0.43	0.48	0.96	0.96	0.57	0.92	0.73	**0.74**
RA-AP	0.67	0.53	0.41	1.00	1.00	0.57	0.93	0.71	**0.73**
Infomap	0.55	0.69	0.47	1.00	1.00	0.52	0.92	0.52	0.71
Louvain	0.46	0.55	0.41	1.00	0.96	0.50	0.93	0.64	0.68
J-AP	0.67	0.50	0.44	1.00	0.96	0.37	0.89	0.34	0.65
ESP-AP	0.57	0.38	0.37	0.96	0.96	0.52	0.92	0.47	0.64
CN-AP	0.11	0.40	0.52	0.90	0.30	0.48	0.88	0.45	0.51
SP-AP	0.73	0.43	0.23	0.68	0.15	0.45	0.63	0.29	0.45

Table 1. The table reports the Normalized Mutual Information (NMI) computed between the ground truth communities and the ones detected by every community detection algorithm for 8 real networks. NMI = 1 indicates a perfect match between the two partitions of the nodes. The methods are ranked by mean performance over the dataset. The table contains also some statistics for each network: number of nodes N, number of edges E, clustering coefficient C, power law degree distribution exponent γ, half of average degree m and number of ground truth communities Nc. The networks are ordered for increasing size N and ties are solved considering the number of edges E.

Method	**Karate**	**Opsahl 8**	**Opsahl 9**	**Opsahl 10**	**Opsahl 11**	**Polbooks**	**Football**	**Polblogs**	**Mean**
EBC-AP	0.75	0.47	0.40	0.98	0.92	0.56	0.89	0.68	**0.71**
RA-AP	0.64	0.53	0.42	1.00	0.96	0.55	0.91	0.57	**0.70**
Infomap	0.54	0.55	0.49	1.00	0.96	0.50	0.92	0.51	0.68
Louvain	0.49	0.51	0.42	1.00	0.96	0.49	0.90	0.63	0.68

Table 2. For each real network, 100 perturbed networks have been generated removing at random the 10% of links. The table reports the Normalized Mutual Information (NMI) computed between the ground truth communities and the ones detected by every community detection algorithm for the 8 real networks, averaged over the 100 iterations. NMI = 1 indicates a perfect match between the two partitions of the nodes. The methods are ranked by mean performance over the dataset.

Method	**Karate**	**Opsahl 8**	**Opsahl 9**	**Opsahl 10**	**Opsahl 11**	**Polbooks**	**Football**	**Polblogs**	**Mean**
EBC-AP	0.59	0.48	0.41	0.97	0.93	0.51	0.89	0.53	**0.66**
RA-AP	0.60	0.51	0.42	0.98	0.95	0.56	0.91	0.35	**0.66**
Louvain	0.45	0.51	0.42	0.98	0.96	0.49	0.90	0.41	0.64
Infomap	0.53	0.55	0.00	0.98	0.00	0.50	0.92	0.31	0.47

Table 3. For each real network, 100 perturbed networks have been generated adding at random the 10% of links. The table reports the Normalized Mutual Information (NMI) computed between the ground truth communities and the ones detected by every community detection algorithm for the 8 real networks, averaged over the 100 iterations. NMI = 1 indicates a perfect match between the two partitions of the nodes. The methods are ranked by mean performance over the dataset.

NetGloVe: Learning Node Representations for Community Detection

Kumaran Gunasekaran[1*], Jeyavaishnavi Muralikumar[1*], Sudarshan Srinivasa Ramanujam[1*], Balasubramaniam Srinivasan[1*], and Fragkiskos D. Malliaros[1,2]

[1] Department of Computer Science and Engineering, UC San Diego, USA
[2] Center for Visual Computing, CentraleSupélec and Inria, France
{kugunase, jmuralik, sus046, bsriniva, fmalliaros}@eng.ucsd.edu

1 Introduction and Problem Statement

Community detection is a fundamental task in network analysis, with plenty of applications in social networking, biology and neuroscience. In the related literature, a variety of algorithms and methodologies have been proposed to identify the community structure of networks, including graph partitioning methods, hierarchical graph clustering, modularity optimization and spectral techniques (such as spectral clustering and modularity optimization).

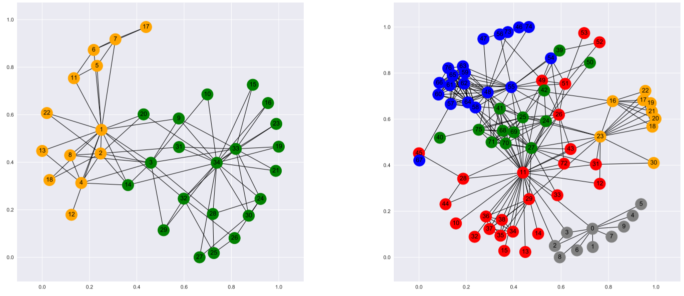
The recent advances in *representation learning* techniques, have allowed us to represent graphs (or nodes) as vectors in a lower dimensional space, that can further be used in graph mining and learning tasks. That way, instead of "manually" extracting features that can be utilized by a graph learning algorithm, we can *learn* informative and discriminative feature representations by solving an optimization problem that takes into account the structural properties of the graph. To this direction, several network feature learning algorithms have been proposed, including node2vec [1] and LINE [4].

The goal of this work is to propose NetGloVe, a new representation learning method for graphs inspired from the domain of Natural Language Processing (NLP), and to examine its application to the task of community detection.

2 Feature Learning with NetGloVe

In this paper, we propose NetGloVe, a node representation learning method inspired by GloVe (Global Vectors for Word Representation) [3], a word embeddings technique in NLP. GloVe uses a log bilinear model to derive vector representations of words, taking into consideration both the word co-occurrence statistics as well as the words context. GloVe is comparable, if not superior, to the Skip-gram model, which considers only the words local context to derive the word representation. Our goal here is to extend GloVe to the context of graphs, by finding a suitable analogy for the word-word co-occurrence matrix used by GloVe. Our intuition is that, nodes that belong to the same communities would have similar embeddings, and thus would be clustered together. That way, we have employed the inverse of the shortest path distance between individual pairs of

[*] Equal contribution; the authors are listed in alphabetical order.



(a) Zachary's karate club (b) Les Misérables

Fig. 1. Community detection results using the NetGloVe embeddings for (a) Zachary's karate club and (b) Les Misérables networks.

nodes in the graph, to populate GloVe's co-occurrence matrix – as we wanted to weight higher nodes that are at close proximity. Thus, the loss function of the node embedding scheme is derived as follows:

$$J = \sum_{i} \sum_{j|d_{ij}<k} f\left(\frac{1}{d_{ij}}\right)\left(w_i^T w_j - \frac{1}{d_{ij}}\right)^2, \tag{1}$$

where f corresponds to a weighting function, w_i and w_j correspond to node vectors and d_{ij} corresponds to the distance between nodes i and j.

3 Experimental Results and Discussion

In our preliminary empirical analysis, we evaluate NetGloVe in the task of community. In particular, we apply NetGloVe to learn feature vectors (i.e., node embeddings) for the nodes of a network, and then we perform k-means clustering to extract the underlying communities. For demonstration purposes, we have applied NetGloVe to the well-known *Zachary's karate club* and *Les Misérables* networks, and the results are depicted in Fig. 1.

Our main experimental evaluation results for NetGloVe are based on artificial networks produced by the LFR benchmark generator [2], where we observe the performance of different methods for a range of mixing parameters. We have used several state-of-the-art baseline methods, including popular representation learning methods (node2vec [1] and LINE [4]), as well as more traditional community detection algorithms (Louvain modularity optimization and spectral clustering). The performance is measured using the normalized mutual information (NMI) criterion between the ground-truth community structure given by the benchmarks and the actual clustering results produced by NetGloVe and the rest baseline methods.

Figure 2 depicts the NMI score obtained by NetGloVe and the baseline methods, on artificial networks with $1,000$ nodes, average degree equal to 20 and maximum degree

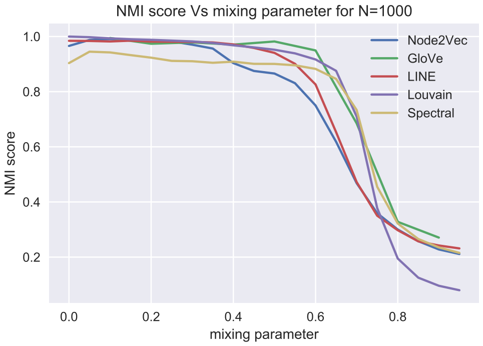


Fig. 2. Comparison of community detection methods for LFR graphs with average degree equal to 20 and maximum degree 50.

equal to 50 (average over multiple realizations of the benchmark graphs). For all the representation learning methods (NetGloVe, node2vec and LINE), we set the number of dimensions to $d = 64$ (i.e., each node is represented as a vector $v \in \mathbb{R}^d$). Lastly, the mixing parameter μ of the generator (x-axis), signifies the ratio of external to internal edges in the graph, with respect to community structure. We observe that NMI declines steeply for all methods for values of μ greater than 0.6, as the community structure becomes less well-defined. Moreover, we can see the NetGloVe performs as good as the rest representation learning methods (Node2vec and LINE) for lower values of μ, and better for higher values of mixing parameter μ – which makes it suitable for community detection in graphs with not well-defined community structure.

Future Work. As future work, we plan to evaluate the performance of NetGloVe on real-world networks with ground-truth community structure. Another future research direction of particular importance is to examine the performance of the features produced by NetGloVe to *supervised* learning tasks over graphs, including the ones of link prediction and node classification.

References

1. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 855–864 (2016)
2. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E, 78(4):046110 (2008)
3. Pennington, J., Socher, R., Manning, C. D.: Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), volume 14, pages 1532–1543 (2014)
4. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. Proceedings of the 24th International Conference on World Wide Web (WWW), pages 1067–1077 (2015)

Methods to reveal communities without the property of ”picking up junk” [*]

Alexander Chepovskiy

National Research University The Higher School of Economics, Moscow, 101000, Russia
aachepovsky@hse.ru,
WWW home page: http://www.hse.ru/staff/aach

1 Introduction

One of the tasks related to the study of the of complex networks is the problem of revealing communities structure. Algorithms to split all vertices into groups, named communities, so that the vertices of each group are more closely related to each other than to the rest of the graph are widely investigated nowadays. This is so-called problem of community detection in graphs of interacting objects. But the most common algorithms to reveal communities has a property of ”picking up junk”. Depending on the initial network analysis purposes, that can be either successful or unsuccessful partition.

In this paper the term ”web of interrelated objects” implies a graph of users that was obtained from the open API of various social networks. Let graphs vertices correspond to the user accounts, while edges — depending on a social network — represent either a ”friendship” relation, then it is a non-oriented graph, or a ”follower” relation, which makes the graph oriented. We consider some features and possible modifications of the popular algorithm for detecting communities — the Blondel algorithm, and present a combined algorithm for detecting intersecting and nested communities in graphs of interacting objects. For the last one Blondel algorithm and another one, based on the search for k-cliques, are taken.

2 Features of Blondel algorithm

A popular method for detecting communities is the Blondel algorithm[1], based on the maximization of Newman-Girvan modularity[2], which is defined as follows:

$$Q = \sum_{i,j} [\frac{A_{ij}}{2m} - \frac{d_i d_j}{4m^2}] \delta(C_i, C_j), \tag{1}$$

Here A is the graph adjacency matrix; d_i is the degree of the vertex i, and m is the number of edges in the graph; C_i — the community that contains the vertex i; δ is the delta function.

The algorithm is agglomerative, because it combines current vertices to new super-vertices step by step in a such way, so that the increase of modularity must be maximal.

[*] Supported by Russian Foundation for Basic Research grants N16-29-09546 and N16-07-00641

This loop breaks when we cannot find a way to increase the modularity. As can be seen from the description of the algorithm, it is necessary to calculate the change of modularity for each vertex when you transfer it from its community to the community of each of the rest neighbour vertices. The modularity change in this case consists of two components: the change that comes from the removal of vertex i from its own community C_i (in this case we associate the vertex with the new community C_k, which contains only this vertex) and the change from adding the vertex i to the new community C_j.

Since the Blondel algorithm optimizes the modularity functional, the results of applying this algorithm are related to certain properties of modularity. Let us consider the expression for the modularity change that comes from transferring the vertex i from the trivial community, that consists only of this vertex, to the community C:

$$\Delta Q = \frac{k_i^C}{m} - \frac{\Sigma_{tot}^C \cdot d_i}{2m^2} \tag{2}$$

where k_i^C is the sum of the edge weights of the edges that are incident to the vertex i and to the community C; m is the sum of the edge weights of the graph; Σ_{tot}^C is the sum of the degrees of the vertices that belong to the community C; d_i is the degree of the vertex i. We obtain the following criterion for transferring a vertex from a trivial community to a community C:

$$k_i^C - \frac{\Sigma_{tot}^C \cdot d_i}{2m} > 0 \tag{3}$$

Moreover, a partition that maximizes the modularity functional on a simple connected graph does not contain trivial communities. It turns out that the leaf vertex in the graph is always connected with its only neighbour community, since otherwise this vertex would lie in a trivial community. And the presence of trivial communities decreases the value of modularity. When we consider two related communities with degrees of each being less than $\sqrt{2m}$, their union will increase the modularity. It is the so-called resolution limit of modularity [5] — in a non-weighted, non-oriented graph partition that has maximal modularity, there can not exist two related communities with degrees of each being less than $\sqrt{2m}$. One of the possible solutions for detecting small communities is the modularity parametrization.

In addition to the properties above, there is another important characteristic considered in [6]. It is when we choose which of the already formed communities we want to add a vertex to, the algorithm is implicitly based on the total weight of the edges that are incident to the vertices of these communities. Therefore it is not uncommon that the algorithm does not add a vertex that has the highest degree to the community with which it has the most number of common edges. It follows that the vertices that are adjacent to a large number of leaves, which unite with them in one community due to the absence of trivial communities, are added to the communities with a small total weight. These properties lead to a partition, which can be called ”junk box”. Due to the properties above, it is reasonable to consider some modifications of the Blondel algorithm that allow to ”pick up junk” less.

Due to the modularity resolution limit, the highest level of the hierarchy of the Blondel algorithm partition does not contain communities that are smaller than a certain size, even if the latter are poorly connected with the rest of the network. However, community data can be detected at lower levels of the hierarchy. So in Fig. 1 it can be seen that a separate community marked by a blue circle, on the second level of the hierarchy, unites with the central community (large community at the bottom of the figure).

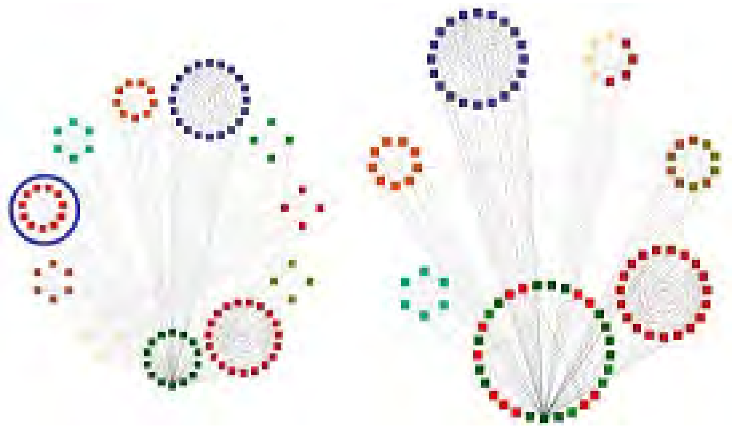


Fig. 1. The first (left) and second (right) levels of the Blondel hierarchy

Another technique, which purpose is to reduce the influence of ”picking up junk” when working with subgraphs, is to consider the degrees of vertices in the original graph instead of the degrees of the vertices of the subgraph. To test this modification, subgraphs of the generated LFR models were obtained by breadth-first search from the random vertex. Comparison of two different approaches — with considering real degrees of vertices and without — shows a significant advantage of the considered modification. For LFR-graphs with number of vertices greater than 4000 the NMI (Normalized Mutual Information) measure of expected partition and that one without considering real degrees is getting less than 0.7, while NMI of expected partition and that one with real degrees consideration is higher than 0.85.

Another method is to use the CPM algorithm [7] after the Blondel one. However, if Blondel’s method identifies a finite community that does not split into parts in terms of a real network, then it is no longer useful to run CPM on it. Therefore, we apply CPM algorithm to each separate C community identified by Blondel method only if the percentage of vertices with high connection centrality does not exceed a predetermined value that depends on the size of the community C. As a result of using CPM, vertices that are not included in cliques are cut off, and there is a problem of assigning them to the community. The solution is a system of two new parameters i_opt and m_opt of the combined method. Depending on their values, not only the cut off vertices are redistributed, but also small communities. This allows to partially influence the resulting structure, and in some cases — to avoid combining graph sheets or small communities into large meaningless communities. In Fig. 2 the Blondel result on the graph G of real network with 158 vertices and 1352 edges is shown.

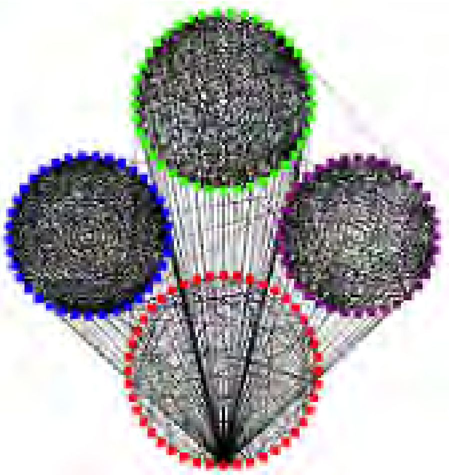


Fig. 2. Communities of G graph obtained by the Blondel algorithm

After applying combined method steps on the graph G, the four communities are transformed into 28 as follows. Within the three meaningful dense communities, minimal changes occur. Algorithm found vertices less connected to the others in them. The fourth community includes leafs of this ego-graph G, and combined method identified 4 communities and 14 individual communities.

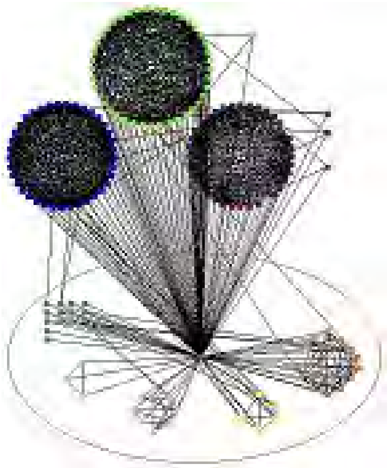


Fig. 3. Communities of G graph obtained by the combined method

3 Results

In this paper authors examined one of the popular hierarchical agglomerative algorithms for detecting communities — the Blondel algorithm. Some properties of this algorithm and the partition of the graph into communities obtained after its application have been described in detail. Moreover, qualitative flaws, which can be critical depending on the initial purpose of network analysis, have been indicated. Possible modifications that can help to obtain other representative results are proposed. In particular, the real degrees of vertices in the subgraph are considered. These variations were tested not only on the generated graphs of the LFR model, but also on real data obtained from social networks. Another approach for revealing intersecting and nested communities, called ”combined method” is also proposed. It allows you not to ”pick up junk” due to algorithm parameters that are responsible for working with small communities and distribution of the cut off vertices.

Summary.

References

1. Blondel V., Guillaume J., Lambiotte R., Lefebvre E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 10. P10008. 12 p.(2008)
2. Girvan M., Newman M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA. Vol.99. 12. P. 7821–7826. (2002)
3. Lancichinetti A., Fortunato S.: Benchmark graphs for testing community detection algorithms. Physical Review. E 78, 046110. (2008)
4. Newman M.E.J., Girvan M. Finding and evaluating community structure in networks // Physical Review. E 69. 026113. 16 p. (2004)
5. Fortunato S., Barthlemy M. Resolution limit in community detection // Proc. Natl. Acad. Sci. USA. Vol. 104. N 36. (2007)
6. Orlov A.O., Chepovskiy A.A. Osobennosti algoritma Blondelya pri vyyavlenii soobshchestv v grafe sotsial’noy seti // V kn .: Trudy Mezhdunarodnoy nauchnoy konferentsii Moskovskogo fiziko-tekhnicheskogo instituta (gosudarstvennyy universitet) i Instituta fiziko-tekhnicheskoy informatiki (SCVRT1516). M., Protvino: Institut fiziko-tekhnicheskoy informatiki. S. 124-129. (2016)
7. Palla G., Dernyi I., Farkas I., Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, vol. 435, no 7043, pp. 814-818. (2005)

Stochastic Local Community Detection in Networks

Yang Li and Hadi Papei

University of Minnesota Duluth, Duluth, MN 55812, USA
yangli@d.umn.edu, hadi.papei@gmail.com,
WWW home page: http://umn.edu/home/yangli

1 Methods

The community structure plays an important role in the dynamics and evolution of complex networks [3]. Loosely speaking, a community is qualitatively defined as the group of vertices whose interconnected edges are much denser than the edges connecting vertices outside of the group. Sometimes, it is desirable to find out the *local* community starting from a given vertex without knowing the global structure. Some algorithms exist [1] [2] [4] [5] whose outputs are deterministic subsets of vertices, which means whether a vertex is in another vertex's local community is binary.

Two vertices can have dissimilar connection strength to another vertex even though they are both in its local community. In social networks, there is a difference between acquaintances and close friends. To assess it quantitatively, we propose an iterative *stochastic* agglomerative procedure to search for local community of a starting vertex v_0. Suppose the local community in the current step is $\mathscr{C}$ with neighboring set $\mathscr{U}$ [2]. The boundary set $\mathscr{B}$ includes those vertices in $\mathscr{C}$ that have at least one neighbor in $\mathscr{U}$. See Figure 1. The local modularity of $\mathscr{C}$ is defined as [2]

$$R = \frac{\text{the number of edges with both ends in } \mathscr{C}}{\text{the number of edges with one or more ends in } \mathscr{C}},$$

which indicates the strength of bounding in $\mathscr{C}$. At each iteration, a new vertex could be added to the community or an vertex currently already in the community could be removed. For each member v_j in $\mathscr{U}$, we compute the change ΔR_j^A if v_j is added to $\mathscr{C}$. Meanwhile, we also compute the change ΔR_k^R if each vertex v_k in $\mathscr{C}$ (except v_0) is removed from $\mathscr{C}$. Let $\Delta R = \{\Delta R^A, \Delta R^R\}$ be the changes of all candidate vertices. Instead of choosing the one with the largest increase ΔR, we select a vertex in a probabilistic way. First, a tolerance parameter $\varepsilon > 0$ is introduced which specifies the maximum allowed reduction in R between iterations. All vertices with $\Delta R < -\varepsilon$ (points below the dashed line in Figure 2) will not be considered. We randomly select one from the remaining vertices with probability $(\Delta R_j + \varepsilon)/\sum_k(\Delta R_k + \varepsilon)$. Vertex with the largest increase will have the largest chance to be selected, but other vertices are not completely eliminated, even though their chances are lower. The iterations keep running until no vertices could be added or removed from $\mathscr{C}$, or R varies little for a certain number of iterations. The subset $\mathscr{C}$ at that moment is then a realization of the local community of v_0.

The procedure above will be repeated for a number of times to get different realizations of v_0's local community. For each vertex v_j in the network, we compute the

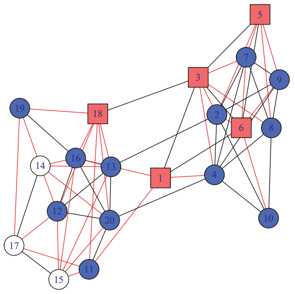


Fig. 1. A synthetic network. The starting vertex is 1. The current local community $\mathscr{C}$ includes the red squares and the neighboring set $\mathscr{U}$ are the blue circles. The boundary $\mathscr{B}$ is equal to $\mathscr{C}$ in this case.

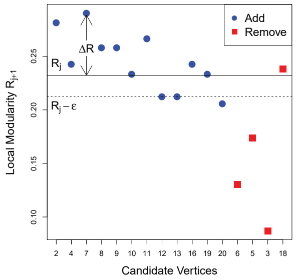


Fig. 2. Changes ΔR for the colored vertices in Figure 1. Blue and red vertices are candidates to be added and removed, respectively. All vertices below the dashed line are not considered. One vertex will be chosen stochastically as explained in the context.

proportion of times that it is included in those realizations, denoted by $P_{v_o v_j}$, the *inclusion probability* of v_j in the community of v_0, whose value indicates how likely v_j is connected to v_0 through its local community structure.

In a deterministic community detection algorithm, if two vertices have the same ΔR, a random tie breaker must come into play. As result of that, different communities may be detected subsequently. It is not the case in our algorithm since the selection is done stochastically. Additionally, parameter ε controls the tightness of the local community. A small ε will generally result in a local community containing vertices with high inclusion probabilities. A larger value of ε, on the other hand, allows the algorithm to probe a larger fraction of the network. More vertices will be included in the local community, and more structures of the network can be revealed.

2 Results

Figure 3 shows a synthetic network with known underlying community structure. It was generated by merging two dense subnetworks $\{1,2,\ldots,20\}$ and $\{21,22,\ldots,40\}$ using a few interconnected edges. We start from vertex 1. After simulating 100 realizations, we get the inclusion probabilities which are shown by the darkness of vertex colors in Figure 3 and also in Figure 4. It is shown that the community structure can be recovered nicely by the algorithm. The inclusion probabilities give us extra information on how close a given vertex is connected to the starting point.

The method is also applied to the famous Zachary's karate club data [6]. Vertex 1 (the instructor) is the starting point and its local community is shown in Figures 5. There is a strong similarity between the actual fission of the club Figure 5(a) and the output from the algorithm Figure 5(b). A cutoff probability of 0.9 gives us exactly the real split of the club into two smaller groups. An interesting point is that vertices 9 and 10 have intermediate probabilities, which lie between the two main structures and are often misclassified by community detection methods. The local community starting from vertex 9 contains most of the vertices, all with similar probabilities.

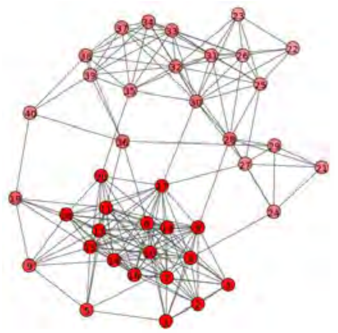


Fig. 3. A synthetic network with known community structure. The darkness of color represents the inclusion probability of being in the local community of vertex 1 . Darker means higher probability.

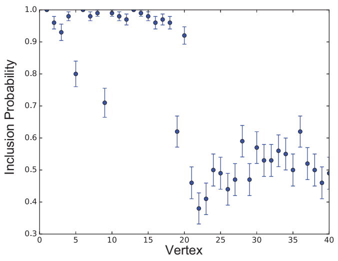


Fig. 4. Plot of inclusion probabilities from Figure 3.

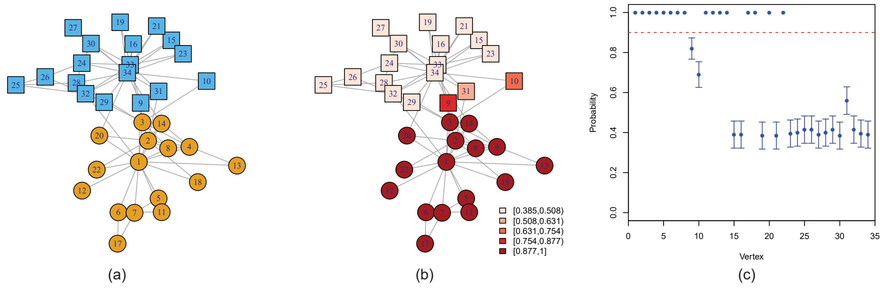


Fig. 5. (a) The karate club network with the actual split into two smaller groups. (b) Color represents probability of a given vertex to be in the local community of vertex 1. (c) Plot of inclusion probabilities. Vertical axis represents the probability a given vertice is in the local community of vertices 1 (the starting vertex).

References

1. Bagrow, J.P., Bollt, E.M.: Local method for detecting communities. Physical Review E 72, 046108 (2005)
2. Clauset, A.: Finding local community structure in networks. Physical Review E 72(2), 026132 (2005)
3. Fortunato, S.: Community detection in graphs. Physics Reports 486(3), 75–174 (2010)
4. Papadopoulos, S., Skusa, A., Vakali, A., Kompatsiaris, Y., Wagner, N.: Bridge bounding: A local approach for efficient community discovery in complex networks. arXiv preprint arXiv:0902.0871 (2009)
5. Rodrigues, F.A., Travieso, G., Costa, L.d.F.: Fast community identification by hierarchical growth. International Journal of Modern Physics C 18, 937–947 (2007)
6. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33(4), 452–473 (1977)

Detecting Dynamic Communities in Social Networks Using Viterbi and Evolutionary Algorithms

Amenah Dahim* and Richard Everson

Department of Computer Science, University of Exeter, UK
* adaa202@exeter.ac.uk

1 Introduction

Finding communities of connected individuals in social networks is important for understanding our society and interactions. Recently attention has turned to discovering the dynamics of communities that change over time. In contrast to existing methods, which treat this as a coupled optimisation problem or post-process communities detected from single snapshots (see e.g., [1]), we formulate the problem in a hidden Markov model (HMM) framework, which allows the most likely sequence of communities to be found using the Viterbi algorithm.

2 Hidden Markov Model and Evolutionary Algorithms

At each time step $t = 1, \dots, T$ we model the nodes N in an graph G^t as belonging to unobserved communities C^t_i, $i = 1, \dots, K_t$. Observations comprise the links (edges) between some of the nodes, so that the entire observation at time t is captured by the adjacency matrix $\mathbf{A}_t$ of the graph G^t.

Our model for the transition probability between states encodes the belief that transitions between similar states are more likely than those between dissimilar states; that is, the network tends to evolve slowly making small transitions. Let $\mathbf{c}_t$ be the N-dimensional vector specifying to which community each node belongs at time t. Then we model the probability of a transition from $\mathbf{c}_{t-1}$ to $\mathbf{c}_t$ as:

$$p(\mathbf{c}_t \,|\, \mathbf{c}_{t-1}) \propto \exp\{\gamma \, \mathrm{NMI}(\mathbf{c}_t, \mathbf{c}_{t-1})\}, \tag{1}$$

where $\mathrm{NMI}(\mathbf{c}_t, \mathbf{c}_{t-1})$ is the Normalised Mutual Information [2], which is commonly used to compare the similarity of cluster or community configurations and $\gamma > 0$ parametrises the strength of the temporal coupling.

The emission probability of observing a particular adjacency matrix models how well a particular community structure $\mathbf{c}_t$ fits the observed adjacency matrix $\mathbf{A}_t$. For example, the modularity $Q(\mathbf{c})$ is a popular measure for evaluating community structure [5] and the emission probability is modeled as:

$$P(\mathbf{A}_t \,|\, \mathbf{c}_t) \propto \exp\{\beta \, Q(\mathbf{c}_t)\}. \tag{2}$$

where $\beta > 0$ sets the discriminative power of the emission probability.

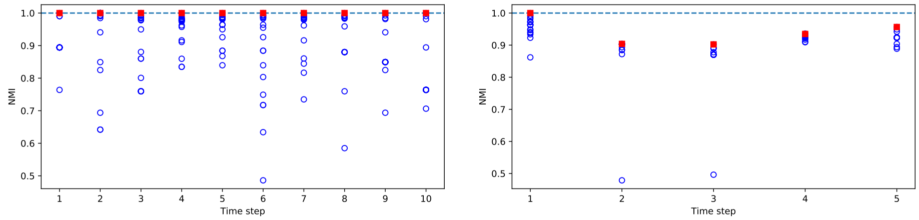


Fig. 1. Synthetic networks. NMI between candidate partitions located by the MOEA and the true partition at each timestep (blue circles). NMI between the true partitions and the Viterbi optimal path are shown as red squares. Left: *SYN-VAR-3*. Right: *SYN-BD*.

The straightforward application of HMM-based filtering and smoothing algorithms is hampered by the vast number of possible states—partitions of nodes into communities —for any real network. To combat this we use a multi-objective evolutionary algorithm to locate a small number of probable states at each time step. Within the space of these probable states we then use the Viterbi algorithm to find the most probable sequence of states, that is the most probable sequence of communities.

In order to find probable candidate states we simultaneously optimise two objectives as functions of the community structure $\mathbf{c}_t$. Communities are characterised by dense connections within each community and sparse connections between them. The first objective (the intra-score) therefore quantifies the density of links within communities, while the second objective (the inter-score) measures the inter-community sparsity. We use the MOEA/D decomposition algorithm [8] to locate an approximation to the Pareto front, the optimal trade-off set between the two objectives. This algorithm is able to locate a wide range of network partitions that are close to the true partition, as measured by the NMI between a partition and the true partition.

The Viterbi dynamic programming algorithm locates the sequence of states/partitions $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T$ that maximise the probability $p(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T \,|\, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_T)$ [6]. Importantly, the algorithm does not require the constants of proportionality in (1) and (2).

3 Results

Figure 1 shows illustrative results on two synthetic networks for which the true partions are known. The networks in Figures 1a (*SYN-VAR-3*, [4]) and 1b (*SYN-BD*, [3]) comprise 256 and 1000 nodes respectively, with community structures that evolve over 10 and 5 time steps. As the figure shows, the MOEA at each timestep has located candidate partitions, some of which are quite distant from the true partition $\mathbf{c}_t^\star$ as measured by the NMI between $\mathbf{c}_t^\star$ and the candidate partition (blue circles). However, the Viterbi algorithm has successfully identified a sequence of partitions that are close to the true partition (red squares). In Figure 1a the located sequence is identical to the true sequence as indicated by the NMI scores of 1. Candidate partitions are more difficult to locate for the *SYN-BD* "birth-death" network, because many nodes are "weak" in the sense that they have a high proportion of connections to nodes outside their community rather

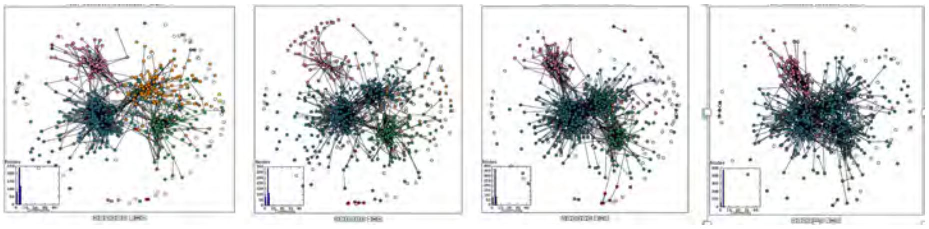


Fig. 2. Weekly MP Twitter communities 16-05-2016 and 16-06-2016. Nodes represent MPs and edges indicate Twitter interactions. Colours indicate distinct communities: blue: remain Labour; pink: Scottish Nationalist Party; orange: remain Conservative; green: leave Conservative.

than intra-community connections. Nonetheless, the Viterbi algorithm has located the sequence of partitions that are closest to the true evolving community structure.

We applied our algorithm to the evolving network Twitter connections between UK Members of Parliament (MPs) in the 85 consecutive weeks from December 2014 to August 2016, the period including a general election and the Brexit referendum held in June 2016 [7]. This network consists of 648 nodes corresponding to the MPs and a link between nodes is made when one MP names another in at least one tweet that week. Figure 2 shows a visualisation of the Viterbi path communities for the four weeks immediately preceding the Brexit referendum. Initially four main communities may be identified: Labour Party remainers, the Scottish Nationalist Party and together with the remain and leave factions of the Conservative Party. As the referendum approaches, the Conservative and Labour remain groups merge, eventually forming a single Twitter conversation including leave MPs.

Acknowledgement: We are grateful to the authors of [7] for the MP Twitter data.

References

1. Aynaud, T., Fleury, E., Guillaume, J., Wang, Q.: Communities in evolving networks: definitions, detection, and analysis techniques. In: Dynamics On and Of Complex Networks, vol. 2, pp. 159–200. Birkhauser, New York (2013)
2. Danon, L., Diaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment 2005(9) (2005)
3. Greene, D., Doyle, D., Cunningham, P.: Tracking the evolution of communities in dynamic social networks. In: Advances in social networks analysis and mining. pp. 176–183 (2010)
4. Kim, M.S., Han, J.: A particle-and-density based evolutionary clustering method for dynamic networks. Proceedings of the VLDB Endowment 2(1), 622–633 (2009)
5. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(2) (2004)
6. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
7. Weaver, I., Williams, H., Cioroianu, I., Williams, M., Coan, T., Banducci, S.: Dynamic social media affiliations among UK politicians (2017), (submitted)
8. Zhang, Q., Li, H.: MOEA/D: A multiobjective evolutionary algorithm based on decomposition. IEEE Transactions on Evolutionary Computation 11(6), 712–731 (2007)

Spectral Multi-scale Community Detection in Temporal Networks with an Application

Zhana Kuncheva[1,2] and Giovanni Montana[2,3]

[1] Clinical Development and Mathematics, C4X Discovery Ltd, M1 3LD, UK.
[2] Department of Mathematics, Imperial College London, SW7 2AZ, UK.
z.kuncheva12@imperial.ac.uk
[3] Department of Biomedical Engineering, King's College London, SE1 7EH, UK.
giovanni.montana@kcl.ac.uk

1 Introduction

The analysis of temporal networks [4] has a wide area of applications in a world of technological advances. An important aspect of temporal network analysis is the discovery of community structures [8]. Real data networks are often very large and the communities are observed to have a hierarchical structure referred to as multi-scale communities. Changes in the community structure over time might take place either at one scale or across all scales of the community structure.

The multilayer formulation [5] of the *modularity maximization* (MM) method [8] introduced in [7] captures the changing multi-scale community structure of temporal networks. This method introduces a coupling between communities in neighboring time layers by allowing inter-layer connections, while different values of the resolution parameter γ enable the detection of multi-scale communities. However, the range of parameter values γ must be manually selected. When dealing with real life data, communities at one or more scales can go undiscovered if appropriate parameter ranges are not selected.

Our recent work on multi-scale community detection in temporal networks [6] proposes a novel Temporal Multi-scale Community Detection (TMSCD) method, which overcomes the obstacles mentioned above. This is achieved by using the spectral properties of the temporal network represented as a multilayer network. In this framework we select automatically the range of relevant scales within which multi-scale community partitions are sought.

2 The Method of Temporal Multi-scale Community Detection (TMSCD)

The proposed TMSCD method relies upon the notion of spectral graph wavelets [3], and is a multilayer extension of the multi-scale community detection procedure via spectral graph wavelets developed in [10].

First, we build a multilayer representation of the temporal network considering new inter-layer weights connecting nodes in neighboring time layers. This introduces dependence between neighboring time points. Second, we apply the definition of a spectral graph wavelet [3] at every node for each time layer. The main idea of the spectral graph wavelet approach is that wavelets at small scales span the local neighborhood of nodes, while wavelets at larger scales span an increasing number of neighboring time layers.

The most essential part in [3] is the design of a wavelet filter function g. The crux of the TMSCD method is the construction of a B-spline based wavelet filter function g adapted for multi-scale community detection in temporal networks. By arguments from Perturbation theory [1] we consider the separate layers as disconnected components, while the inter-layer weights as perturbations. In this way, when studying the spectral properties of the supra-Laplacian (the Laplacian of the multilayer formulation), we take into account the fundamental difference between within-layer and inter-layer edges.

By Perturbation theory the eigenvectors corresponding to the smallest eigenvalues of the supra-Laplacian are linear combinations of the eigenvectors (corresponding to the 0 eigenvalues) of the Laplacian matrices of the separate time layers. From spectral graph theory [2], it is known that an eigenvector corresponding to the 0 eigenvalue of the Laplacian matrix is not informative of the community structure. For this reason, the eigenvectors of the supra-Laplacian matrix, which can be obtained as approximations of these eigenvectors, cannot be used to identify within-layer communities. Hence, we propose a procedure for the selection of certain larger eigenvalues, which are "informative" of the prevalent community structure over time. Thus we reconsider the role of the Fiedler vector in community detection for temporal networks.

The contribution of the eigenvectors at different scales s is controlled by a B-spline based wavelet filter function g, whose few parameters are automatically selected. The filter puts more weight to smaller "informative" eigenvalues when large scale communities are sought, and more weight to larger eigenvalues when small scale communities are sought.

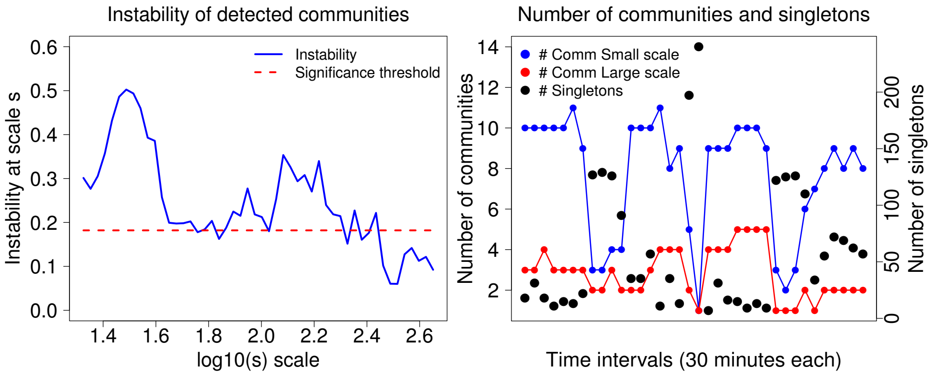
A series of simulations on various benchmarks presented in [6] show the competitive performance of TMSCD to MM. These simulations show that the proposed inter-layer weights perform better than fixed inter-layer weights. Another advantage of TMSCD over MM is the automated selection of scales' ranges at which multi-scale communities should be sought.

3 Application to Primary School Data

Here we present for the first time initial results from applying the TMSCD method to the temporal social patterns appearing in a primary school [9]. Data on face-to-face interactions between 242 students and teachers from 10 classes were collected for two days, which we subdivide into 36 intervals of 30 minutes. For each interval, the network of interactions has a link between two individuals if those individuals had at least one contact during the corresponding interval.

After applying the TMSCD method to this temporal network, we assess the statistical significance of the communities at each scale, which identified two stable scales of partitions, Fig. 1(a). For the number of communities at each time point for the two visualized partitions and the number of singletons see Fig. 1(b).

For the smaller scale, consecutive time intervals which overlap with pre- and after-lunch class periods group students in around $9-10$ communities. This means that each class is in its own community and communication takes place between class peers.



(a) Instability of detected communities at each scale s with significance threshold.

(b) Number of communities and singletons at each time point for two different scales.

Fig. 1. Results from applying TMSCD to temporal social patterns appearing in a primary school.

Fewer defined communities appear during lunch and at the beginning/end of the school day when many students walk around and no long face-to-face contact is observed. The large scale captures the split between lower grades (first, second and third) and higher grades (fourth, fifth and sixth), or between lower (first and second), middle (third and fourth) and higher (fifth and sixth) grades. Overall, these results demonstrate the strength of TMSCD to detect multi-scale communities in temporal networks. Future work involves more thorough real life applications and improvements of the scalability issues of the method. The author ZK acknowledges partial support by grant no. I 02/19 of Bulgarian NSF.

References

1. Bhatia, R.: Matrix Analysis, Graduate Texts in Mathematics, vol. 169. Springer New York, New York, NY (1997)
2. Chung, F.: Spectral Graph Theory. CBMS (1996)
3. Hammond, D.K., Vandergheynst, P., Gribonval, R.: Wavelets on Graphs via Spectral Graph Theory. Appl. Comput. Harmon. Anal. 30(2), 129–150 (mar 2011)
4. Holme, P., Saramäki, J.: Temporal Networks. Phys. Rep. 519(3), 97–125 (oct 2012)
5. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer Networks. Multilayer Networks 2(3), 203–271 (2014)
6. Kuncheva, Z., Montana, G.: Multi-scale Community Detection in Temporal Networks Using Spectral Graph Wavelets (aug 2017)
7. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. Science (80-.). 328 (2010)
8. Newman, M.E.J.: Modularity and Community Structure in Networks. Proc. Natl. Acad. Sci. U. S. A. 103(23), 8577–82 (jun 2006)
9. Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.F., Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., Vanhems, P.: High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. PLoS One 6(8), e23176 (aug 2011)
10. Tremblay, N., Borgnat, P.: Graph Wavelets for Multiscale Community Mining. IEEE Trans. Signal Process. 62(20), 5227–5239 (oct 2014)

A comparison of hierarchical community detection algorithms

Zhao Yang[1], Juan I. Perotti[2], and Claudio J. Tessone[1,2]

[1] URPP Social Networks, University of Zurich, Andreasstrasse 15,
CH-8050 Zürich, Switzerland
[2] IMT School for Advanced Studies Lucca, Piazza San Francesco 19,
I-55100 Lucca, Italy

In this study, for the first time we perform an extensive analysis of state-of-the-art hierarchical community detection algorithms. Three hierarchical community detection algorithms have been chosen, which are: Infomap [1], a recursive application of Louvain method for the generation of hierarchies [2] [3] and the Minimum Description Length implementation of the Hierarchical Stochastic Block Model (HSBM)[4]. To do so, we introduce a novel benchmark graph with ground truth which preserves several properties of real-world networks with hierarchical community structure. It is able to generate networks that have (i) an arbitrary number of hierarchical levels; (ii) a power-law degree distribution; (iii) a power-law community size distribution; and (iv) the smallest or at least reasonable sizes possible. We argue that most of the existing hierarchical benchmark models can not generate graphs with power-law community size distribution. To quantify the performance of the above mentioned algorithms, we resort the recently introduced hierarchical mutual information metric [3]. This allows us quantification quality of the algorithms in recovering the hierarchical community structure. We also demonstrate that these benchmark graphs exhibit a variety of topological transitions between co-existing ground truths.

In order to construct the benchmark graph (termed RB-LFR) we start from a standard non-hierarchical LFR benchmark network, which we consider as the seed network motif for an adapted Ravasz-Barabási procedure for constructing hierarchical networks [5][6]. We have used the latter because it has clear defined hierarchical structure and it can be extended to the multiple levels. The networks produce two different well-defined ground truths, and the mixing parameter controls a topological transition between them. Taking the two-level RB-LFR benchmark graphs as an example, when the mixing parameter of the seed LFR benchmark is small, its community structure and that of its replicas are well-defined. On the first level, the RB-LFR benchmark displays as many communities as the seed LFR has, i.e. C communities. Each community in this first layer contains one community of the seed LFR together with all its replicas. At the second level, each community of the first one contains $R+1$ sub-communities – one for each replica plus the seed one – summing a total of $C \times (R+1)$ sub-communities in the complete network. When the mixing parameter μ is increased, the community structure of the seed LFR becomes more fuzzy and harder to detect. This also happens to the seed and all replicas communities with the RB-LFR benchmark. This happens while the number of inter-layer links remain the same regardless of μ. Therefore, the seed LFR and the replicas may be interpreted as $R+1$ communities at the first layer. Each of them has as many sub-communities at the second level as the seed LFR had, i.e. C.

Again, the total number of sub-communities at the second level is $(R+1)\times C$ but, this time, such a number is reached through different means. If the mixing parameter of the seed LFR becomes too large, then the communities become impossible to detect and the community structure of the RB-LFR benchmark network has a single level.

The results of our tests are shown in Fig. 1. In the left panels the accuracy of the different community detection algorithms are quantified by the average value of the NHMI (i.e. *Normalized Hierarchical Mutual Information*) computed between the detected hierarchical community structures and the different ground truths. In the center column, the similarity is quantified with the average NMI (i.e. *Normalized Mutual Information*) computed between the detected partitions at the second level and those exhibited by the different ground truths. In the right panels, the similarity is quantified by the difference HMI - MI between the HMI computed for the full hierarchies and the MI computed for the partitions at the first level. The tested methods are Infomap, Louvain, and HSBM from top to bottom. Taking the top-left panel as an example: Infomap can unveil the community structure until $\mu \approx 0.6$. For $\mu < 0.1$, it detects the first type of ground truth, and for $0.2 < \mu < 0.6$, it detects the second type of ground truth. We observe a clear transition between the ground truths; in both regions, the NHMI reaches values close to 1 making apparent that the algorithm gives a description of the hierarchy very close to the ground truth. For large μ, Infomap detects a flat community structure. This result showcases that the RB-LFR benchmark shows a clear hierarchical community structure which can be recognized successfully by Infomap. Comparing panels (a) to (d), and (g) of Fig. 1, we observe that the new benchmark poses a challenging task that can test the performance of the algorithms, and that the capabilities of the different algorithms depend non-trivially on the network topology. We note here that the poor performance of the HSBM is most likely related to its approach, i.e. a bottom-up approach, while the other two methods are taking the top-down approaches to build the hierarchies. Finally, the difference HMI - MI is shown in Fig. 1c, f, & i. It gives the contribution that the second level has on the HMI. In other words, it quantifies how accurately the algorithms detect the second level and how relevant is the corresponding contribution as measured by the HMI. It is clear that such contribution is non-negligible, showing the convenience of *HMI* as a measure for the comparison of hierarchical community structures, when compared to the traditional *MI*.

To conclude, we have found that the newly introduced RB-LFR benchmark graphs pose challenging tests to state-of-the-art hierarchical community detection algorithms. More specifically, the tests on the two-level RB-LFR benchmark graphs indicate that Infomap outperforms the other two methods in terms of accuracy. Our benchmark graphs, while parsimonious, exhibit a rich phenomenology including a variety of topological transitions between co-existing ground truths. Additionally, our tests have also validated that the recently introduced *Hierarchical Mutual Information (HMI)* suits better for the comparison of hierarchical partitions than the traditional *Mutual Information (MI)* does.

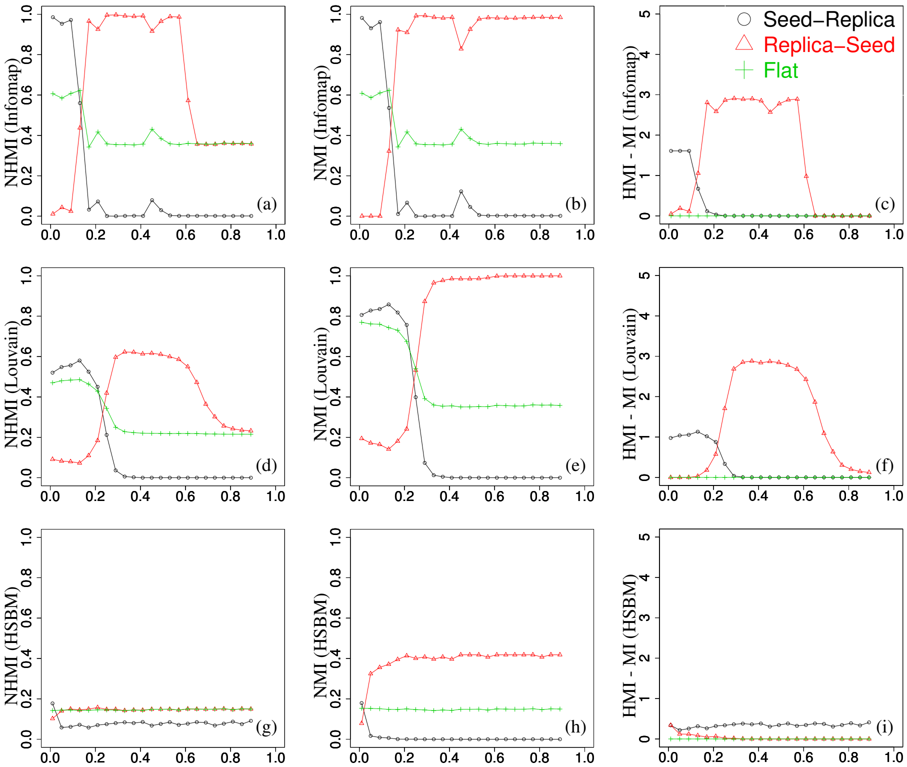


Fig. 1. Average NHMI, NMI, and (HMI - MI) as a function of the mixing parameter, μ at the left, middle and right panels, respectively. From top to bottom, the methods are Infomap, Louvain, and HSBM. Averages are computed over 10 different network realizations with the same set of parameters of the seed LFR benchmark.

References

1. Rosvall, M., Bergstrom, C. T.: Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. 105, 1118–1123 (2008)
2. Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E. : Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. 2008, P10008 (2008)
3. Perotti, J. I., Tessone, C. J., Caldarelli, G.: Hierarchical mutual information for the comparison of hierarchical community structures in complex networks. Phys. Rev. E 92, 062825 (2015)
4. Peixoto, T. P.: Hierarchical block structures and high-resolution model selection in large networks. Phys. Rev. X 4, 011047 (2014)
5. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E 78, 046101 (2008)
6. Ravasz, E., Barabási, A. L.: Hierarchical organization in complex networks. Phys. Rev. E 67, 026112 (2003)

Part XI

Resilience and Control

The effective structure of complex networks: Canalization in the dynamics of complex networks drives dynamics, criticality and control

Rion Brattig Correia[1,2], Alexander Gates[1,3], Santosh Manicka[1], Manuel Marques-Pita[1,2], Xuan Wang[1], and Luis M. Rocha[1,2,3*]

1 School of Informatics, Computing & Engineering, Indiana University, Bloomington, IN, USA.
2 Instituto Gulbenkian de Ciencia, Oeiras, Portugal.
3 Program in Cognitive Science, Indiana, University, Bloomington, IN, USA.
* Correspondence should be addressed to L.M.R. (email: rocha@indiana.edu)

Abstract:

Network Science has provided predictive models of many complex systems from molecular biology to social interactions. Most of this success is achieved by reducing multivariate dynamics to a graph of static interactions. Such network structure approach has provided many insights about the organization of complex systems. However, there is also a need to understand how to *control* them; for example, to revert a diseased cell to a healthy state in systems biology models of biochemical regulation. Based on recent work [1,2] we show that the control of complex networks crucially depends on *redundancy* that exists at the level of variable dynamics. To understand the effect of such redundancy, we study automata networks−both systems biology models and large random ensembles of Boolean networks (BN). In these discrete dynamical systems, redundancy is conceptualized as *canalization*: when a subset of inputs is sufficient to determine the output of an automaton. We discuss two types of canalization: *effective connectivity* and *input symmetry* [2].

First, we show that effective connectivity strongly influences the controllability of multivariate dynamics. Indeed, predictions made by structure-only methods can both undershoot and overshoot the number and which sets of variables actually control BN. Specifically, we discuss the effect of effective connectivity on several structure-only controllability theories: structural controllability (SC), minimum dominating sets (MDS), and feedback vertex sets (FVS) [1,3]. While SC , MDS and similar theories assume linear dynamics, their application to real-World networks that are not expected to be linear is widespread from systems biomedicine to network neuroscience. Therefore, using BN to study their ability to predict control is very reasonable, as BN are one of the simplest nonlinear dynamical systems known; if prediction is poor for BN (in both realistic biochemical regulation models and theoretical ensembles [1]) there is no reason to think prediction would be better for real-World networks that are very likely to be nonlinear too.

To understand how control and information effectively propagate in such complex systems, we uncover the *effective graph* that results after computation of effective connectivity. To study the effect of input symmetry, we further develop our *dynamics canalization map*, a parsimonious dynamical system representation of the original BN obtained after removal of all redundancy [2]. Mapping canalization in BN via these representations allows us to understand how control pathways operate, aiding the discovery of dynamical modularity [4] and robustness present in such systems [2]. We also demonstrate that effective connectivity is a tuning parameter of BN dynamics [5], leading to a new theory for criticality, which significantly outperforms the existing theory in predicting the dynamical regime of BN (chaos or order). Input symmetry is also shown to affect criticality, especially in networks with large in-degree. Moreover, we argue that the two forms of canalization characterize qualitatively distinct

phenomena, since Boolean functions cover the space of both measures and prediction performance of criticality is optimized for models which parameterize the two forms separately [6]. Finally, we will showcase a new Python toolbox that allows the computation of all canalization measures, as well as the effective graph and the dynamics canalization map. We will demonstrate it by computing the canalization of a battery 50+ systems biology automata networks.

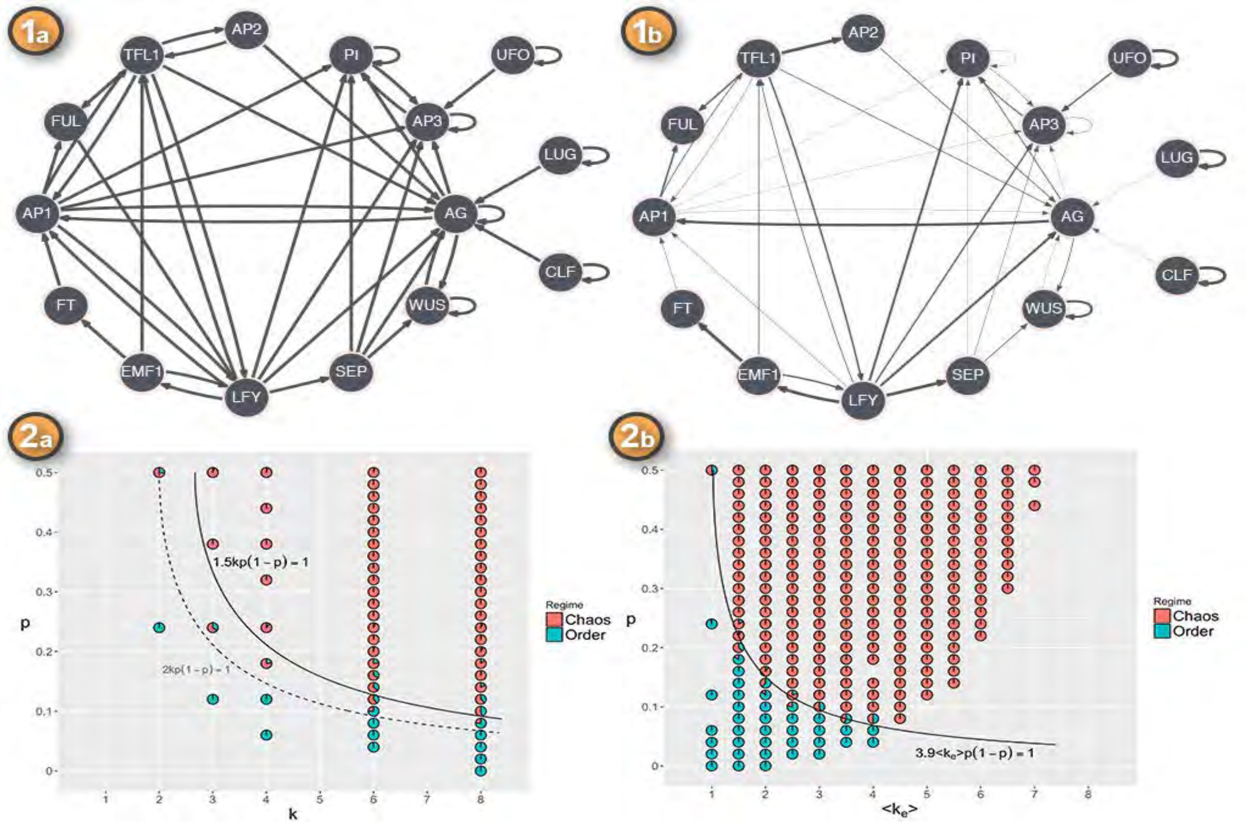


Figure 1: (1a) *Arabidopsis Thaliana* gene regulatory BN model [7] of cell-fate determination during floral organ specification; nodes denote genes, directed edges indicate that source gene regulates expression of target gene [3]. (1b) Effective graph of Thaliana BN shown in 1.a; most edges are much less effective in regulation than original model predicts, with various edges being completely (FUL → LFY, AP1 → LFY, AP2 → TFL1) or almost completely redundant (AG → AG, SEP → AG; both with less than 1% effectivity) in controlling the network, which is not taken into account by structural controllability theories leading to erroneous predictions about controllability [1,3]. (2a) Phase diagram in the k-p (in-degree per bias) space, showing the dynamical regimes of ensembles of BN samples; green pie slices denote the proportion of BN with ordered dynamics, and red indicates proportion of BN with chaotic dynamics; critical boundaries displayed for best model in this space (bold) and existing theory of criticality (dashed) [5]. (2b) Phase diagram in the ke-p (effective connectivity per bias) space; critical boundary displayed for best model in this space (bold) [5]; the best model using effective connectivity (2.b) significantly outperforms the classification performance of best model in space displayed in 2a, with more than a 40% performance increase [5], which can be clearly grasped by visually comparing misclassifications in 2.a and 2b.

References

[1] A. Gates and L.M. Rocha. [2016]. Scientific Reports 6, 24456.

[2] M. Marques-Pita and L.M.Rocha [2013]. PLOS One, 8(3): e55946.

[3] A. Gates, R.B. Correia and L.M. Rocha. [2017]. In Preparation.

[4] A. Kolchinsky, A. Gates and L.M. Rocha. [2015] Phys. Rev. E. 92, 060801(R).

[5] M. Marques-Pita, S. Manicka and L.M.Rocha. [2017]. In Preparation.

[6] S. Manicka and L.M.Rocha. [2017]. In Preparation.

[7] Espinosa-Soto,C., Padilla-Longoria, P. & Alvarez-Buylla, E. R.. *The Plant Cell Online* **16**, 2923–2939 (2004).

Improving Coordination in Heterogeneous Human-Agent Complex Networks: The case of Vertex-Covering Problem

Pouria Babvey, Babak Heydari

Stevens Institute of Technology, Castle Point on Hudson, Hoboken, NJ 07030, USA
{pbabvey, babak.heydari}@stevens.edu

1 Introduction

Many complex systems of future consist of networks of interacting humans and autonomous agents. This heterogeneous set of agents often come together to collectively perform a set of complex distributed tasks. Performance of such networks depends strongly on the level of efficient coordination among constituting agents [1]. Having autonomous technology agents enables designers of such networks to use them in order to steer coordination behavior of the network as a whole, by manipulating the strategic behavior of human agents.

In this study, we use this lens to establish new roles for robots as social mediators between people to improve collective performance of heterogeneous networks in solving a decentralized complex problem (i.e., Vertex-Covering Problem). For the constituent agents, we consider randomness and patience levels as two defining dimensions of heterogeneity for a network. The patience level models the extent by which agents balance long-term higher efficiency and short-term rapid improvement in performance. Randomness provides a cascade of benefits by creating more exploring opportunities for the network as a whole [2]. Through agent-based simulation, we study the role of randomness and patience level in shaping coordination, and using these two parameters in order to improve the coordination efficiency of the network as a whole.

2 Method

To implement this study, we selected the Vertex Covering Problem, one of the benchmark complex problems in computer science. A vertex cover of a network is a set of nodes such that each edge in the network is incident to at least one vertex of the set. A large number of algorithms have been suggested for solving this NP-complete problem; but for the sake of this study, we need to select an algorithm that is decentralized in nature in order to make it suitable for a collective task. The algorithm also needs to be compatible with the heuristics by which human agents approach solving the problem.

To satisfy these two conditions, we created a new decentralized heuristic algorithm. Unlike most similar problems in which agents are placed on network nodes, we considered agents to be placed on the edges, coordinating with each other to choose a set of covering nodes.

More precisely, in an iterative process, each edge in the network chooses one of its endpoints as a covering node. A node is exempted from the set of covering nodes if it is not selected by any of its incident edges. As such, coordination becomes critical since edges incident on the same node need to coordinate with each other to exempt their common endpoint.

To assess the performance of the proposed method we used the relative metric with an exponential penalty for error. We consider $C_\pi = e^{\left[\frac{V}{V^*}-1\right]}$ in which V^* denotes the minimum number of covering nodes, and V shows the number of covering nodes the method suggests. Efficiency is inversely proportional to C_π, with 1 representing the maximum efficiency. We performed an agent-based simulation based on the proposed algorithm on a variety of network structures. Our results show a Goldilocks effect for both of randomness and patience in the coordination process. For high randomness levels, agents explore the solution space almost steadily and the changes could not be followed up by the neighbors. For low randomness, edges do not explore the solution space enough and may trap in sub-optimal solutions. As for the effect of patience, higher patience restrains the agents from reaching a consensus since each agent waits for other agents to comply with its strategy; while for lower patience, deficient coordinations linger on for a long time.

We further tested the effect of heterogeneity of randomness and patience on coordination. For this, we simulated the process for different sample groups. The results show that in case we distribute the agents randomly, heterogeneity is not beneficial for coordination. In the next step, we developed allocation policies based on agents intrinsic parameters and network structure. We used some network metrics like edge-betweenness and edge-centrality to establish agent distribution policy.

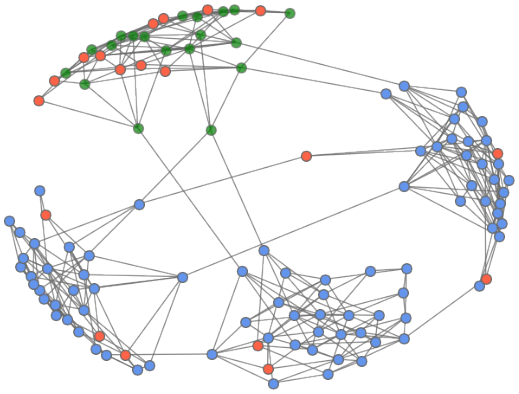
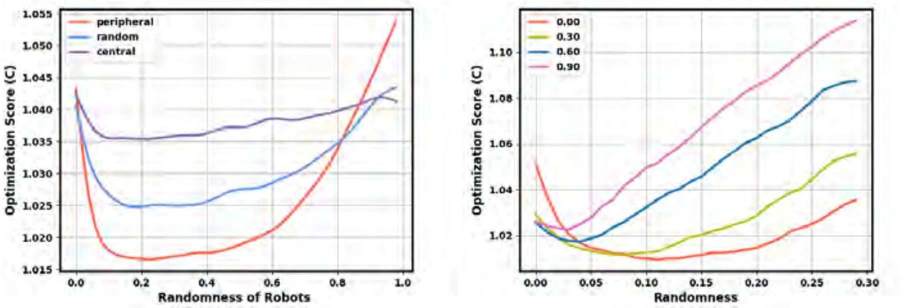


Fig. 1. The result of the process for a modular network; Red nodes has been exempted. Distribution policy has been applied for the module with green nodes.

The results indicate that with the same average parameters using the locating policy heterogeneous community outperforms the homogeneous one. In the next step, we established a policy to improve the overall performance using robots. To replicate humans behavior, we used model-free learning as a popular class of Reinforcement Learning algorithms. *Model-free* directly estimates the state-action values based on the historical returns [3]. Assuming that humans use a myopic policy with a low randomness, we injected some noise to the system by locating robots with exceedingly high randomness, as 10-15% of all community members in the network.

The robots were fabricated with varying levels of behavioral randomness. We also used three different policies for locating robots on the network. In the first policy, we randomly assigned robots to edges. In the two other policies, we selected edges incident on the nodes with lower and higher centralities, respectively. The results showed that in all three cases adding robots boosts the performance; however, for the case where we place the robots on peripheral edges, improvement is more significant.



(a) (b)

Fig. 2. The average performance of 10000 random networks with 40 nodes (a) The results of using fabricated robots with three different allocation policy for robots (b) The results of different overall impatience levels; Each impatience level favors a different randomness level

3 Summary of Results

The proposed method shows the power of coordination among myopic autonomous agents to solve computationally infeasible problems. Specifically, this work demonstrates two key results; first of all, we showed that there are an optimal randomness and patience level for any network structure. This means that as the designer of such networks, we can boost the system-level performance using robots with a carefully fabricated level of randomness and patience level. Secondly, even though heterogeneity in patience and randomness is not beneficial in general, the performance can be improved by choosing the network location of agents based on their types. Although these results are for a specific complex task, we can expect that the obtained results could be generalized to a wide variety of coordination problems in heterogeneous networks.

References

1. Kearns, M., Suri, S., Montfort, N.: Human Subject Networks An Experimental Study of the Coloring Problem on An Experimental Study of the Coloring Problem on Human Subject Networks. 824, (2006).
2. Shirado, H., Christakis, N.A.: Locally noisy autonomous agents improve global human coordination in network experiments. Nature. 545, 370374 (2017).
3. Renaudo, E., Girard, B., Chatila, R., Khamassi, M.: Respective advantages and disadvantages of model-based and model-free reinforcement learning in a robotics neuro-inspired cognitive architecture. (2015).

Key Factors and Mechanisms of Sustainability - Network Analysis Comparision of Texts and Causal-loop Diagrams

Gyula Dorgo, Gergely Honti, Daniel Leitold, and Janos Abonyi

MTA - PE Lendulet Monitoring Complex Systems Research Group, University of Pannonia,
POB. 158, Veszprem, Hungary
janos@abonyilab.com,
WWW home page: http://www.abonyilab.com

Abstract. To identify the key factors of sustainability, we study local, regional and global ecological models and the related scientific papers and strategical documents. We extract networks from causal-loop diagrams of ecological models and generate graph-based representations of texts to evaluate the similarity of different knowledge representations, identify the most relevant variables and study their effect on the controllability and observability of ecosystems. In a case study, we explore the similarity of the clusters of the extracted networks to the analyse the importance of water-related variables.

Because the variables in a complex ecosystem are strongly interdependent, it is hard to identify which inputs contribute to an observed output, and what is the real importance of a variable in an ecosystem. Our goal is to provide a better understanding of these interactions by the simultaneous structural analysis of dynamical models and the related texts of sustainability.

We identify state variables that have a central role in controllability and observability of state space models identified from causal loop diagrams. To perform controllability and observability-specific analysis and evaluate measures of node centrality we developed a NOCAD MATLAB toolbox [1].

We enrich this analysis with expert knowledge represented by graphs extracted from scientific papers and strategical documents of sustainability and look for similar patterns in these networks. To demonstrate the applicability of the proposed approach, we highlight the effect of the water-related state variables in complex ecosystems and show how these are discussed in the Sustainable development goals of the United Nations (see Figure 1.).

Using our methods, we can determine the most relevant state variables and the critical intervention and measurement points. We study well-known dynamical models of water resources (like ANEMI, shown in Figure 2.) to demonstrate the applicability of the methodology. Water-related state variables were identified as important parts of every model. Results from specialized, local models highlighted that the water quality is the most critical problem.

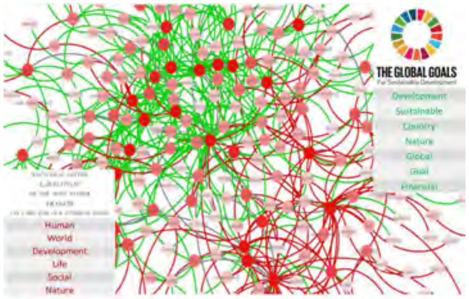


Fig. 1. We extract networks from the bag of words models of scientific papers, sustainability reports and strategic documents. The figure illustrates the similarity of the Sustainable development goals of the United Nations and the Laudato Si encyclical letter. The extracted keywords demonstrate how the viewpoints of UN and Pope differs.

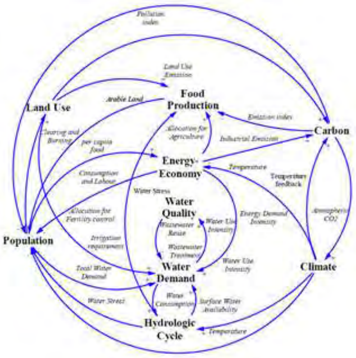


Fig. 2. Causal-loop diagram of the ANEMI model of water resources that we transform into state-space model to analyse the controlability, observability, and importance the variables.

The developed methodology can be used to determine the most relevant state variables and critical intervention and measurement points. The comparative analysis of models and the related texts provides suggestions for model improvement. We hope that the simultaneous analysis of causal loop models and expert knowledge represented by texts can open new possibilities in decision-making.

Acknowledgements We acknowledge the financial support of EFOP-3.6.1-16-2016-00015 Smart Specialization Startegy (S3) - Comprehensive Institutional Development Program at the University of Pannonia. The work of Gyula Dorgo was suppoerted throughthe New National Excellence Program of the Ministry of Human Capacities.

References

1. Leitold, D., Vathy-Fogarassy, Á., Abonyi, J.: Controllability and observability in complex networks–the effect of connection types. Scientific Reports 7(1), 151 (2017)

Impact of removing nodes on the controllability of complex networks

Stylianos Savvopoulos and Sotiris Moschoyiannis

Department of Computer Science, University of Surrey, GU7 XH, UK
`S.Savvopoulos@surrey.ac.uk`, `S.Moschoyiannis@surrey.ac.uk`,
WWW home page:
`https://www.surrey.ac.uk/cs/people/sotiris_moschoyiannis`

1 Introduction

Complexity theory has been used to study a wide range of systems in biology and nature but also business and socio-technical systems, e.g., see [2]. The ultimate objective is to develop the capability of steering a complex system towards a desired outcome. Recent developments in network controllability [3] concerning the reworking of the problem of finding *minimal control configurations* allow the use of the polynomial time Hopcroft-Karp algorithm instead of exponential time solutions. Subsequent approaches build on this result to determine the precise control nodes, or d̑rivers, in each minimal control configuration [6], [1]. A browser-based analytical tool, *CCTool*[1], for identifying such drivers automatically in a complex network has been developed in [5].

One key characteristic of a complex system is that it continuously evolves, e.g., due to dynamic changes in the roles, states and behaviours of the entities involved. This means that in addition to determining driver nodes it is appropriate to consider an evolving topology of the underlying complex network, and investigate the effect of removing nodes (and edges) on the corresponding minimal control configurations. The work presented here focuses on arriving at a classification of the nodes based on the effect their removal has on controllability of the network.

2 Results

Methodology. We consider three categories in terms of cardinality of the maximum matching, C_{MM}: a node is *delete-redundant*, iff C_{MM} is unchanged; *delete-ordinary*, iff C_{MM} is reduced by one; and, *delete-critical* iff C_{MM} is reduced by more than one.

We applied the following adaptation of the algorithm found in [4], [3], [5] for classifying nodes based on the effect their removal has on controllability of the network.

1. Find one maximum matching of the graph $G(V,E)$ by applying Hopcroft-Karp.
   - Convert the graph into a bipartite graph: $G_b\ (V_+, V_-, E)$ where V_+ is the out-set, and V_- the in-set
   - Run the Hopcroft-Karp algorithm on the bigraph G_b to find a maximum matching (denote as MMG) and denote the size of MMG as CG

[1] Available at: http://cctool.herokuapp.com

2. Obtain sets of matched nodes and S-matched nodes of G with respect to MMG
   - A set of matched nodes in V_- is one of matched nodes of G (denote as M_-)
   - A set of matched nodes in V_+ is one of S-matched nodes of G (denote as M_+)
   - Denote a set of matched or S-matched nodes of G as M ($M = M_+ \cup M_-$)
   - Nodes that are not contained in the set M are not S-redundant and are therefore delete-redundant
3. Pick one node in M and identify the category of the node
   - Pick one node n in M and denote the node in V_+ as n_+, and in V_- as n_-
   - Create a subgraph S_b by removing all edges incident with node n_+ and n_-
   - Create a matching MS by removing matched edges incident with nodes n_+, n_-
   - Find a maximum matching of S_b and denote the size as CS
     - Find an augmenting path with regards to MS in subgraph S_b
     - If there is an augmenting path, use it to augment MS
     - Once more, find an augmenting path and augment MS
     - Augmented matching MS is a maximum one of S_b
   - Identify the category of the node n
     - If CS = GS then n is delete redundant
     - If CS = $GS - 1$ then n is delete ordinary
     - If CS = GS - 2 then n is delete critical
4. Repeat step 3 until all nodes in M are identified by using MMG and G_b

where a matched node is the end point of a matched link; an S-matched node is the start point of a matched link; an S-redundant node is a node that is always a matched node or an S-matched node, in all maximum matchings; and, a node is a delete redundant node if and only if the node is not an S-redundant node.

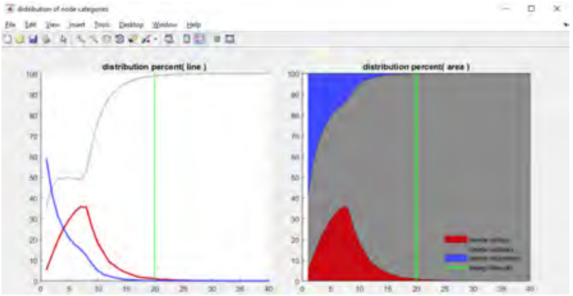
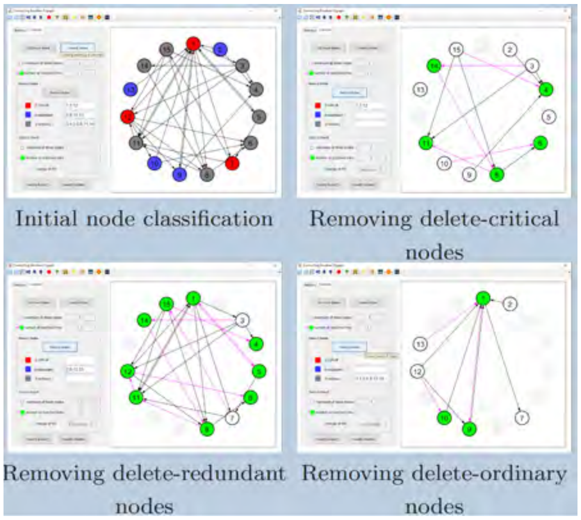


Fig. 1. Distribution of node categories ($N = 1000$)

Summary. Fig.1 shows the distribution of the different categories of nodes considered. If the edge probability increases, all nodes tend to be delete-ordinary, while delete-redundant nodes disappear gradually. When the edge probability exceeds a threshold, the fraction of delete-critical nodes, which was increasing, starts to rapidly decrease.

Fig.2 shows the initial state of the network (top left) and the effect of removing different categories of nodes. Our experiments showed that when 3 delete-critical



Initial node classification

Removing delete-critical nodes

Removing delete-redundant nodes

Removing delete-ordinary nodes

Fig. 2. Node classification based on impact on network controllability

nodes are removed the number of driver nodes increases by 3. When 4 delete-redundant nodes are removed the number of driver nodes decreases by 2 while removing 8 delete-ordinary nodes results in no change in the number of driver nodes.

In terms of robustness of the node categories devised, we note the following results.

When some nodes are removed from the network, all node classifications are stable with delete-redundant nodes being the most unstable. When the edge probability increases, the delete-ordinary nodes are the most stable. When 10% of the nodes are removed from the network, then over 70% of the nodes' categories are not changed. Also, the delete-ordinary category is the most stable one.

References

1. Haghighi, R., Namazi, H.R.: Algorithm for identifying minimum driver nodes based on structural controllability. Mathematical Problems in Engineering 2015 (2015)
2. Krause, P., Razavi, A., Moschoyiannis, S., Marinos, A.: Stability and complexity in digital ecosystems. In: IEEE DEST 2009. pp. 85–90 (2009)
3. Liu, Y.Y., Slotine, J.J., Barabasi, A.L.: Controllability of complex networks. Nature 473, 167–173 (May 2011)
4. Liu, Y.Y., Slotine, J.J., Barabasi, A.L.: Control centrality and hierarchical structure in complex networks. PLoS ONE 7 (2012)
5. Moschoyiannis, S., Elia, N., Penn, A., Lloyd, D.J., Knight, C.: A web-based tool for identifying strategic intervention points in complex systems. In: Games for the Synthesis of Complex Systems (CASSTING'16 @ ETAPS 2016). vol. 220, pp. 39–52. EPTCS (2016)
6. Penn, A.S., Knight, C.J.K., Chalkias, G., Velenturf, A.P.M., Lloyd, D.J.B.: Extending participatory fuzzy cognitive mapping with a control nodes methodology: A case study of the development of a bio-based economy in the humber region, uk. In: S. Gray, M.P., Jordan, R. (eds.) Environmental Modeling with Stakeholders. Springer (2016)

Measuring the stability of complex hierarchical networks

Maryam Zamani[1] and Tamas Vicsek[1]

Department of Biological Physics, Eotvos University, Pazmany Peter setany 1/A 1117,
Budapest, Hungary,
zamani@caesar.elte.hu,
WWW home page: https://sites.google.com/site/maryamzamaniphysics/

Abstract. We investigate the stability of networks that emerge from simulations optimizing an efficiency function which can be related to both the behaviour of organizations and to the Hamiltonian of spin-glasses. Using this quantitative approach we find a number of expected and highly non-trivial results for the obtained locally optimal networks, stability increases with growing efficiency, the same perturbation results in a larger change for more efficient states, due to the huge number of possible optimal states only a small fraction of them exhibits resilience.

1 Introduction

Stability is one of the most essential features of complex systems ranging from ecological to social, communication and economic networks [1]. Stability of a system can be investigated from several perspectives including the perhaps two most essential ones: resistance and resilience. In Ref. [4] we introduced a model in order to interpret the apparently glassy behaviour of hierarchical organizations and their corresponding network of interactions. The model leads to a complex behaviour of the efficiency function associated with the performance of networked organizations resembling the phenomena displayed by the so called spin glass model [2]. Here we address the question of central importance: how stable is the network structure against perturbations? What is the relation between efficiency and stability? The stability of locally optimal states of directed complex networks is examined by adding perturbations (noise) to the system. While after optimization the structure of the network freezes in one of its locally optimal states, the effect of noise relocates links in the system. Change in efficiency [3] is studied between local and noisy states and are compared for different values of noise. Network resistance against perturbation is investigated by measuring the number of steps that should be taken before disturbing efficient state. Network resilience is considered by looking at the ability of the system to return to its local optimal state after turning off the noise.

2 Method

In Ref. [4], an efficiency function is developed for a typical organization which is built from interacting individuals with different abilities a_i.

$$E_{eff} = \frac{1}{N}\sum_{ij} J_{ij} a_i a_j, \qquad (1)$$

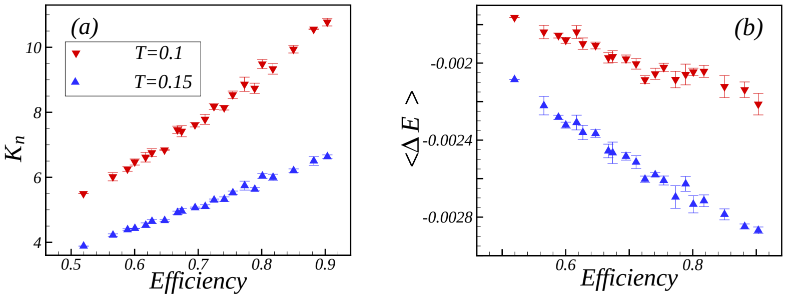


Fig. 1. (a) K_n(the number of steps that noise is turned on until the first change in efficiency occurs) versus efficiency . (b) Average of efficiency difference $\langle \Delta E \rangle$ between optimal and unstable states versus efficiency in two different values of noise (temperature).

where N is the number of nodes. Directed edges between individuals have signs corresponding to their harmonic ($J_{ij} = +1$) or antagonistic ($J_{ij} = -1$) relation, ($J_{ij} = 0$) if the two nodes are not connected. Efficiency function is maximized in order to find local optimal states of the networks using Monte Carlo simulation. Network efficiency and its corresponding structure has a glassy behavior meaning that it does not converge to a unique efficient state. Maximizing the efficiency function leads to complex behavior and hierarchical structures emerge during the process of searching the optimal states. The distribution of local maxima of efficiencies and their corresponding global reaching centrality (*GRC*) or level of hierarchy (method for calculating *GRC* is discussed in Ref. [3]) values indicates that optimal states fall into two categories with high and low *GRC* [4]. The question is how much these local optimal states are stable, from resistance point of view against external perturbation and also their ability to return to their optimal state after turning off the noise (resilience). External perturbation can be a noise in local optimal state of the system, optimal states are achieved through Monte Carlo simulation by randomly relocating the position of the edges in temperature close to zero. In each Monte Carlo step, the efficiency is calculated. If the efficiency is higher than the previous step it is accepted and if it is lower, it is accepted by Boltzmann probability ($\exp(-\frac{\Delta E_{eff}}{T})$). To reach the saturated highly efficient state, temperature (T) in Boltzmann probability should be close to zero. After reaching the optical states where efficiency saturates, we increase temperature to implement the noise which increases the Boltzmann probability. The noise is kept on until efficiency changes. Then the difference between efficiencies in two states (optimal and unstable states) as well as the number of steps taken to see the first change in efficiency are calculated.

3 Results

We start our interpretation of the results in Figure 1-a. by the observation that for a given noise (perturbation) higher values of K_n imply higher stability and vice versa. Networks with higher efficiencies need more steps to deviate from

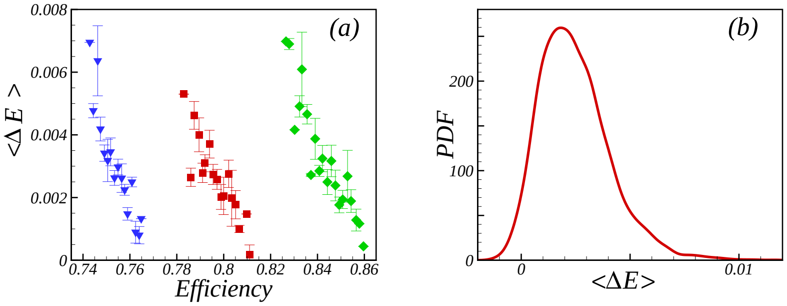


Fig. 2. (a) Average efficiency difference between two local optimal states $\langle \Delta E \rangle$ before turning on the noise and after its turn off versus efficiency, these results were obtained for three different initial full graphs. (b) Probability density function of $\langle \Delta E \rangle$ is skewed with a peak close to zero.

its optimal state and exhibit a higher level of resistance against external perturbation. According to Figure 1-b, for systems with higher efficiency the absolute value of $\langle \Delta E \rangle$ (reaction) of the system is larger in response to a perturbation. Since we already established that higher efficiency translates to higher stability we can conclude based on our findings so far the following: Systems with higher stability are less susceptible to external perturbation but once the perturbation is large enough, they undergo a more pronounced change. The networks resilience is shown in Figure 2. To model this, we start with an optimal state and turn on the noise changing T from nearly zero to $T = 0.1$. After 32 steps, noise is switched off and the system is allowed to recover from its unstable state and converge to a local optimal state again. For larger efficiencies the network returns to the same initial optimal states $\langle \Delta E = 0 \rangle$ after turning off the noise (high resilience). Positive values of $\langle \Delta E \rangle$ demonstrate networks switch to a more efficient state after turning off the noise. Figure 2-b shows probability density function of $\langle \Delta E \rangle$ with positive skew and a high peak close to zero.

Acknowledgement

This research was supported in part by the grant EU FP7 RED-Alert No: 740688

References

1. Dutta, B., Jackson, M.O.: The efficiency and stability of directed communication networks. Review of Economic Design 5(3) (9 2000)
2. Mezard, M., Montanari, A.: Information,Physics, and computation. Oxford University Press, https://books.google.hu/books?id=jhCM7i0a6UUC, 3 edn. (1 2009)
3. Mones, E., Vicsek, L., Vicsek, T.: Hierarchical self-organization of non-cooperating individuals 7(3), e33799 (3 2012)
4. Zamani, M., Vicsek, T.: Glassy nature of hierarchical organizations. Scientific Reports 7(1382) (5 2017)

Topological resilience in non-normal networked systems

Malbor Asllani[1] and Timoteo Carletti[1]

Department of Mathematics & naXys, Namur Institute for Complex Systems, University of Namur, Rue Rempart de la Vierge, 8, B-5000 Namur - Belgium

1 Introduction

The network of interactions in complex systems, strongly influences their resilience, the system capability to resist to external perturbations or structural damages and to promptly recover thereafter [1]. The phenomenon manifests itself in different domains, e.g. cascade failures in computer networks or parasitic species invasion in ecosystems [2]. Understanding the networks topological features that affect the resilience phenomenon remains a challenging goal of the design of robust complex systems [3]. For this purpose, we introduce the concept of non-normal network, namely a network whose adjacency is a non-normal matrix [4] and provide an algorithm for building networks with such property [5]. We prove that the non-normality character of the network of interactions amplifies the response of the system to exogenous disturbances and can drastically change the global dynamics. To illustrate the power of non-normal network, we provide an illustrative application to ecology by proposing a mechanism to mute the Allee effect, the phenomenon according to which for initial low densities the species is not able to survive. This manifestation of unexpected species invasion eventually describes a new theory of patterns formation involving a single diffusing species inspired by a transient instability principle.

2 Results

The non-normal assumption is thus responsible for unexpected outcomes on the dynamics, constituting hence a basis for the resilience of networked systems being related to the structural property of the network. To illustrate the potentiality of this mechanism we describe a major application in ecology, the Allee effect [6], whose dynamics can be described by the following diffusively coupled equations:

$$\frac{dx_i}{dt} = rx_i(1-x_i)\left(\frac{x_i}{A}-1\right)+D\sum_{j=1}^{M}L_{ij}x_j, \forall i, \tag{1}$$

where x_i denotes the species density in the i-th patch, r the reproduction rate, A the Allee coefficient, D the diffusion coefficient and L the Laplacian matrix.

If the network of interactions is normal and the initial conditions do not exceed the Allee threshold, $x_i(0) < A, \forall i$, then the species goes extinct and diffusion cannot prevent it. Conversely if the underlying network belongs to the family of non-normal networks, the system fate turns upside down and the population will survive reducing thus the system resilience. The explanation for this behavior can be found in the competition between the diffusion mechanism and the reproduction rate. It can happen that the initial transient amplification induced by the non-normality is strong enough to surpass the Allee threshold, at least in some of the patches, and consequently the system saturates avoiding the extinction. A schematic illustration of the resilience response of the Allee model on a non-normal network is presented in Fig. 1. This theoretical approach can be used to describe different scenarios in ecology like the introduction of new individuals in a particular habitat for the goal of species conservation or biological control from invasive species [7]. Often, the achievement of such goals is strongly conditioned by the Allee effect.

The previously discussed model where the non-normality facilitates the survival of the species and avoid the Allee effect does in fact, yields a non homogeneous non-linear pattern. One of the mostly diffused mechanism responsible for the pattern formation, is the one introduced by A. Turing in his seminal paper on morphogenesis [8]. According to him in order to be able to realize the celebrated Turing patterns, the minimal requisites are the presence of at least two species, the “activator” (capable to trigger their own growth) and the “inhibitor” (antagonist to the former, impede any further growth once diffusing), and moreover the ratio of their diffusion coefficient (inhibitor vs. activator) should be larger than some threshold. In the new scenario of non-normal networks, this topological feature of the network of interactions will force the inhibitors in some of the nodes, to initially increase their concentration until they saturate in the non-linear phase. What is remarkable is that the species which apparently tends to go extinct because of the negative growth rate, exploit a faster diffusion process that makes the species to spread before the individuals counteract, and eventually lead to self-organization. This is thus a new mechanism, different form the Turing one, capable to explain the pattern formation process, observe moreover that one species is enough to activate this phenomenon, provided it diffuses on a non-normal network.

We have also developed a method to generate non-normal networks based on the Newman-Watts mechanism [9]. More precisely, taking a random weighted directed ring, acting as the backbone, we add long-range links with random weights considerably smaller than the ones of the ring, reflecting thus the observation that the small-world topology is widely spread in Nature: direct interaction between far away nodes, is less probable and weaker with respect to closer ones. The adjacency matrix results thus almost triangular, a structure that closely resembles the Jordan blocks of the canonical form of a non-normal matrix [4].

Summary. We studied the role of non-normal topology in the resilience phenomenon of dynamical systems defined on directed networks. In the case of non-normal networks, unexpectedly, small perturbations will follow an initial amplification that can lead the system to a new state, possibly far from the initial one. Once the non-normality guides the system toward a non-homogeneous equilibrium state, the process at work can be cast in the framework of patterns formation driven by dynamical instability. The single species case hereby presented shows the impact of the network topology on the self-organization process, allowing the formation of patterns beyond the Turing mechanism.

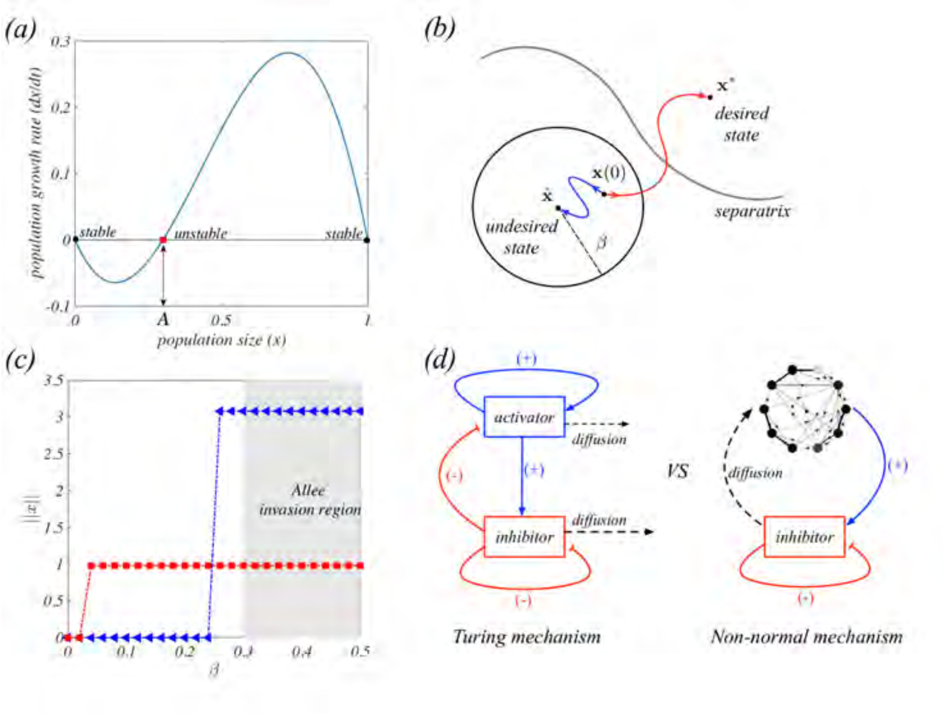


Fig. 1. Resilience and the Allee effect. (a) The Allee effect states that for $x(0) < A$ the species fails to survive. **(b)** Non-normal (red) versus normal (blue) system. **(c)** Bifurcation diagram of the networked Allee model shows that for the non-normal systems (red) the survival zone is strongly increased. **(d)** Turing pattern mechanism versus the non-normal one.

References

1. Neubert M.G., Caswell H.: Alternatives to resilience for measuring the responses of ecological systems to perturbations. Ecology 78(3), 653–665, (1997)
2. Motter A. E., Lai Y.-C.: Cascade-based attacks on complex networks. Phys. Rev. E 66, 065102 (2002)
3. Gao J., Barzel B., Barabási A.-L.: Universal resilience patterns in complex networks. Nature 530, 307–312, (2016)
4. Trefethen L.N., Embree M.: Spectra and pseudospectra: The behavior of nonnormal matrices and operators. Princeton University Press (2005)
5. Asllani M., Carletti T.: Topological resilience in non-normal networked systems. arXiv:1706.02703, (2017)
6. Allee W.C., Bowen E.: Studies in animal aggregations: mass protection against colloidal silver among goldfishes. J. Exp. Zool. 61(2), 185–207 (1932)
7. Courchamp F., Clutton-Brock T., Grenfell B.: Inverse density dependence and the Alle effect. TREE 14(10), 405–410 (1999)
8. Turing, A.: The chemical basis of morphogenesis. Phil. Trans. R. Soc. B 237, 37–72 (1952)
9. Newman M.E.J., Watts D.J.: Scaling and percolation in the small-world network model. Phys. Rev. E 60, 7332–7342 (1999)

Bounding Robustness via Kirchhoff Index

Monica Bianchi, Gian Paolo Clemente, Alessandra Cornaro, and Anna Torriero

Department of Mathematics, Finance and Econometrics,
Catholic University, Milan,
monica.bianchi@unicatt.it, gianpaolo.clemente@unicatt.it,
alessandra.cornaro@unicatt.it, anna.torriero@unicatt.it

1 Introduction

Bounding robustness in complex networks has gained increasing attention in the literature. Network robustness research has indeed been carried out by scientists with different backgrounds, like mathematics, physics, computer science and biology. As a result, quite a lot of different approaches to capture the robustness properties of a network have been undertaken. Traditionally, the concept of robustness was mainly centered on graph connectivity. Recently, a more contemporary definition has been developed. According to [16], it is defined as the ability of a network to maintain its total throughput under node and link removal. Under this definition, the dynamic processes that run over a network must be taken into consideration.
In this framework several robustness metrics based on network topology or spectral graph theory have been developed (see [4], [8], [9], [13]). In particular, we focus on spectral graph theory where robustness is measured by means of functions of eigenvalues of the Laplacian matrix associated to a graph ([10] and [19]). Indeed, this paper is aimed to the inspection of a graph measure called *effective graph resistance*, also known as *Kirchhoff index* (or *resistance distance*), derived from the field of electric circuit analysis ([14]). The Kirchhoff index has undergone intense scrutiny in recent years and a variety of techniques have been used, including graph theory, algebra (the study of the Laplacian and of the normalized Laplacian), electric networks, probabilistic arguments involving hitting times of random walks ([6] and [7]) and discrete potential theory (equilibrium measures and Wiener capacities), among others. It is defined as the accumulated effective resistance between all pairs of vertices. This index is widely used in Mathematical Chemistry, Computational Biology and, more generally in Network Analysis in order to describe the graph topology.
It is worth pointing out that the Kirchhoff index can be highly valuable and informative as a robustness measure of a network, showing the ability of a network to continue performing well when it is subject to failure and/or attack. In fact, the pairwise effective resistance measures the vulnerability of a connection between a pair of vertices that considers both the number of paths between the vertices and their length. A small value of the effective graph resistance therefore indicates a robust network. Several works studied indeed the Kirchhoff index in networks that are topologically changed. For example, Ghosh et al [12] study the minimization of the effective graph resistance by allocating link weights in weighted graphs. Van Mieghem et al in [20] show the relation between the Kirchhoff index and the linear degree correlation coefficient. Abbas

et al in [1] reduce the Kirchhoff index of a graph by adding links in a step-wise way. Finally, [17] focuses on Kirchhoff index as an indicator of robustness in complex networks when single links are added or removed. In particular, being the calculation of this index computationally intensive for large networks, they provide upper and lower bounds when an edge is added or removed. Some of them are based on the relation with the algebraic connectivity ([11]), another measure of network robustness.
In this paper, we discuss a methodology aimed at obtaining some new and tighter bounds of this graph invariant when edges are added or removed which takes advantage of real analysis techniques, based on majorization theory and optimization of functions which preserve the majorization order, the so-called *Schur-convex* functions. One major advantage of this approach is to provide a unified framework for recovering many well-known upper and lower bounds obtained with a variety of methods, as well as providing better ones. It is worth pointing out that the localization of topological indices is typically carried out by applying classical inequalities such as the Cauchy-Schwarz inequality or the arithmetic-geometric-harmonic mean inequalities. Within this topological robustness framework, we propose to use our bounds, obtained by these techniques, for robustness assessment of complex networks. Further research regards a generalization to weighted and/or directed networks and the analysis of the correlation between alternative topological metrics.

2 Preliminaries

Let us first recall some basic concepts from graph theory (for more details we refer the reader to [5] and [18]). In this paper we consider a simple, connected and undirected graph with n nodes and m edges. Let $\pi = (d_1, d_2, .., d_n)$ be the degree sequence of G, where $d_1 \geq d_2 \geq \cdots \geq d_n$, d_i is the degree of vertex v_i and $\sum_{i=1}^{n} d_i = 2m$. $A(G)$ is the adjacency matrix of G and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ is the set of (real) eigenvalues of $A(G)$. The matrix $L(G) = D(G) - A(G)$ is called Laplacian matrix of G, where $D(G)$ is the diagonal matrix of vertex degrees. Let $\mu_1 \geq \mu_2 \geq ... \geq \mu_n$ be the eigenvalues of $L(G)$, where $\sum\limits_{i=1}^{n} \mu_i = 0$ and $\sum\limits_{i=1}^{n} \mu_i^2 = 2m$. We now present some basic facts about majorization theory. The main references about majorization order and Schur convexity are the classical book [15] and the paper [2] for the notations and techniques.

Definition 1. *Given two vectors* $\mathbf{y}$, $\mathbf{z} \in D = \{\mathbf{x} \in \mathbb{R}^n : x_1 \geq x_2 \geq ... \geq x_n\}$*, the majorization order* $\mathbf{y} \trianglelefteq \mathbf{z}$ *means:*

$$\begin{cases} \langle \mathbf{y}, \mathbf{s^k} \rangle \leq \langle \mathbf{z}, \mathbf{s^k} \rangle, \ k = 1, ..., (n-1) \\ \langle \mathbf{y}, \mathbf{s^n} \rangle = \langle \mathbf{z}, \mathbf{s^n} \rangle \end{cases}$$

where $\langle \cdot, \cdot \rangle$ *is the inner product in* $\mathbb{R}^n$ *and* $\mathbf{s^j} = [\underbrace{1, 1, \cdots, 1}_{j}, \underbrace{0, 0, \cdots 0}_{n-j}], \quad j = 1, 2, \cdots, n.$

Given a closed subset $S \subseteq \Sigma_a = D \cap \{\mathbf{x} \in \mathbb{R}^n_+ : \langle \mathbf{x}, \mathbf{s^n} \rangle = a\}$, where a is a positive real number, let us consider the following optimization problem

$$\mathrm{Min}_{\mathbf{x} \in S} \ \phi(\mathbf{x}). \tag{1}$$

If the objective function ϕ is Schur-convex, i.e. $\mathbf{x} \trianglelefteq \mathbf{y}$ implies $\phi(\mathbf{x}) \leq \phi(\mathbf{y})$, and the set S has a minimal element $\mathbf{x}_*(S)$ with respect to the majorization order, then $\mathbf{x}_*(S)$ solves problem (1), that is

$$\phi(\mathbf{x}) \geq \phi(\mathbf{x}_*(S)) \;\; \text{for all } \mathbf{x} \in S.$$

It is worthwhile to notice that if $S' \subseteq S$ the inequality $\mathbf{x}_*(S) \trianglelefteq \mathbf{x}_*(S')$ holds and thus

$$\phi(\mathbf{x}) \geq \phi(\mathbf{x}_*(S')) \geq \phi(\mathbf{x}_*(S)) \; \text{for all } \mathbf{x} \in S'. \tag{2}$$

On the other hand, if the objective function ϕ is Schur-concave, i.e. $-\phi$ is Schur-convex, then

$$\phi(\mathbf{x}) \leq \phi(\mathbf{x}_*(S')) \leq \phi(\mathbf{x}_*(S) \; \text{for all } \mathbf{x} \in S'. \tag{3}$$

A very important class of Schur-convex (Schur-concave) functions can be built adding convex (concave) functions of one variable. Indeed, given an interval $I \subset \mathbb{R}$, and a convex function $g : I \to \mathbb{R}$, the function $\phi(\mathbf{x}) = \sum_{i=1}^{n} g(x_i)$ is Schur-convex on $I^n = \underbrace{I \times I \times \cdots \times I}_{n-times}$. The corresponding result holds if g is concave on I^n.

In [2], the extremal vectors for some closed particular subsets $S \subseteq \Sigma_a$ have been computed.

3 Theoretical and numerical results

We assume to obtain a graph $G'(n, m+h)$ by adding $1 \leq h \leq \dfrac{n(n-1)}{2} - m$ links to the graph $G(n,m)$. By using majorization techniques we are able to find the extremal vectors of some suitable subsets of $\mathbb{R}^n$ that encode information about the spectrum of the Laplacian matrix. In this way, by exploit the fact that the Kirchhoff index is a Schur-convex function we can find the following bound:

$$R(G^{'}) \geq n\left(\frac{1}{d_1+1} + \frac{1}{d_2} + \frac{(n-3)^2}{2m+2h-1-d_1-d_2}\right). \tag{4}$$

In a similar way, we define $G^{''}$ the graph obtained by removing $1 \leq h \leq n-1-m$ links from $G(n,m)$ and we can derive:

$$R(G^{''}) \geq n\left(\frac{1}{d_1+1-2h} + \frac{1}{d_2-2h} + \frac{(n-3)^2}{2m-2h-1-d_1-d_2+4h}\right) \tag{5}$$

We now compare our bounds with those in [17], where the following lower bounds are provided, when $h = 1$:

$$K(G') \geq \frac{K(G)}{1+\dfrac{n\rho}{2}}, \tag{6}$$

where ρ is the diameter of G,

$$K(G'') \geq \frac{n(n-1)^2}{2(m-1)}. \tag{7}$$

In tables 1 and 2 we show that the proposed bounds (4) and (5), evaluated for $h = 1$, are tighter than the bounds (6) and (7).

n	$K(G)$	$K(G')$	Our bound (4)	Bound Wang et al. (6)
10	19.86	18.97	16.53	1.81
20	48.71	47.80	43.57	1.57
30	66.21	65.70	60.80	2.14
40	78.64	78.49	75.63	1.92
50	106.76	106.59	102.06	2.09
100	191.36	191.28	188.38	1.89
200	409.23	409.19	405.21	2.04
500	1001.28	1001.26	997.12	2.00
1000	1999.25	1999.24	1995.26	2.00

Table 1. Comparison between lower bounds of $K(G)$ after one link addition. Graphs are randomly generated by using Erdös-Rényi (ER) model varying the number of vertices n and with probability $p = 0.5$.

n	$K(G)$	$K(G'')$	Our bound (5)	Bound Wang et al. (7)
10	19.86	21.17	17.78	17.61
20	48.71	49.26	44.31	44.02
30	66.21	66.69	61.16	60.94
40	78.64	78.81	75.87	75.67
50	106.76	106.92	102.28	102.08
100	191.36	191.43	188.49	188.41
200	409.23	409.27	405.28	405.21
500	1001.28	1001.29	997.15	997.10
1000	1999.25	1999.26	1995.27	1995.24

Table 2. Comparison between lower bounds of $K(G)$ after one link removal. Graphs are randomly generated by using Erdös-Rényi (ER) model varying the number of vertices n and with probability $p = 0.5$.

Additionally, we show that these bounds are effective also when more than one links is added or removed.

References

1. Abbas, W., Egerstedt, M. (2012), Robust graph topologies for networked systems, in 3rd IFAC Workshop on Distributed Estimation and Control in Networked Systems, pp.85-90.
2. Bianchi M., Cornaro A., Torriero A. (2013) Majorization under constraints and bounds of the second Zagreb index, furthercoming on Mathematical Inequalities and Applications, 16-2, 329-347.
3. Bianchi M., Cornaro A., Torriero A. (2013) A majorization method for localizing graph topological indices , Discrete Applied Mathematics, 161, 2731-2739.
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez and M. Hwanga, D. U. (2006) Complex networks: structure and dynamics. Physics Reports, 424:175–308.
5. Bollobás B., (1990) Graph Theory: An introductory course, Springer Verlag, New York.
6. Broder A. Z., Karlin A. R., (1989) Bounds on the cover time. J. Theoret. Proba- bility 2-1:101–120.
7. Chandra A., Raghavan P., Ruzzo W., Smolensky R., Tiwari P. (1989) The electrical resistance of a graph captures its commute and cover times. STOC, 574–586.
8. Costa, F.. Rodrigues, F.A., Travieso, G., Villas Boas, P.R. (2007) Characterization of complex networks: A survey of measurements. Advances in Physics, 56-1:167–242.
9. Dorogovtsev, S.N., Mendes, J.F.F. (2002) Evolution of networks. Advances in Physics, 51:1079-1187.
10. Ellens, W., Spieksma, F.M. , Van Mieghem, P., Jamakovic, A., Kooij, R.E. (2011), Effective graph resistance, Linear algebra and its applications, 2491–2506.
11. Fiedler, M. (1973) Algebraic connectivity of graphs. Czechoslovak Mathematical Journal, 23:298-305.
12. Ghosh, A., Boyd, S., Saberi, A. (2008) Minimizing Effective Graph Resistance of a Graph, SIAM Rev. 50, 37.
13. Jamakovic, A. (2008) Characterization of complex networks, application to robustness analysis, Phd thesis, Delft University of Technology, 2008.
14. Klein, D. J. and Randic, M. (1993) Resistance Distance, J. Math. Chem, 12, 81.
15. Marshall A. W., Olkin I., Arnold B., (2011) Inequalities: Theory of Majorization and Its Applications, Springer.
16. Sydney, A., Scoglio, C. M., Schumm, P., Kooij, R. E. (2008) Elasticity: Topological characterization of robustness in complex networks, in Proceedings of the 3rd International Conference on Bio-Inspired Models of Network, Information and Computing Sytems, Brussels, Belgium, pp. 19:119:8.
17. Wang X., Pournaras E., Kooij R.E., Van Mieghem P. (2014) Improving robustness of complex networks via the effective graph resistance, European Physical Journal B. 87:221.
18. Wilson R. J., (1996) Introduction to graph theory, Addison Wesley.
19. Van Mieghem, P. (2011) Graph Spectra for Complex Networks, Cambridge University Press, Cambridge, U.K.
20. Van Mieghem, P., Ge, X., Schumm, P., Trajanovski, S., Wang, H. (2010) Spectral graph analysis of modularity and assortativity, Phys. Rev. E 82, 056113.

Hydraulically informed graph theoretic metric for the resilience analysis of water supply networks

Aly-Joy Ulusoy and Ivan Stoianov

InfraSense Labs, Dept. of Civil and Environmental Eng., Imperial College London SW7 2BU, London, UK,
aly-joy.ulusoy15@imperial.ac.uk

Risk assessment is a key aspect of the design and operation of engineering infrastructures. It has led to the development of the 3Rs risk management strategy, which builds upon the notions of network redundancy, resilience and reliability.

Reliability is defined as the study of both the probability and consequences of network component failure. However, as a result of the lack of universally accepted probabilistic approach to component failure, explicit consideration of reliability in water supply network design has proven a challenging issue [2]. To overcome the statistical limitations of reliability measures [6], our work focuses on the consequences of component failure and the capacity of the system to cope with them, which are assessed by the notions of resilience and redundancy.

The redundancy of a network derives from its connectivity properties and its analysis is supported by graph theory. Despite providing valuable insight into the network's structure, graph theoretic metrics do not allow to quantitatively assess the capacity of network paths. Hydraulic resilience analyses, on the other hand, use hydraulic simulation results to measure link capacities and derive the demand a network can meet under failure conditions [6]. In practice, however, hydraulic data is subject to quality and availability restrictions. The computational complexity of repeated hydraulic simulations also grows exponentially with the size of the network, due to the non-linearity of water distribution systems. These limitations can make the use of hydraulic metrics unpractical for the assessment of the resilience of real-life water distribution networks.

This paper introduces a surrogate measure of hydraulic resilience called water flow edge betweenness centrality (WFEBC) that aims to overcome the current limitations of both graph theoretic and hydraulic metrics. Water flow edge betweenness centrality builds upon the latest developments in graph theory, to which it integrates elements of physical characteristics of water supply networks. Inspired by Newman's random walk betweenness centrality measure, WFEBC takes an alternative approach to traditional shortest-path betweenness measures [3], which constrain the flow to the geodesic paths of the network, yielding a more realistic representation of the flow conditions. Due to space constraints, the reader is referred to [5] and [1] for the mathematical definition of current flow betweenness centrality. The main contribution of this work is the introduction of the following modifications to Newman's definition, allowing for its application to the criticality analysis of water distribution systems:

- The adjacency matrix is weighted: the random walks are computed on edges weighted by resistance coefficients derived from the physical laws conditioning flow dynamics in the network.

- Whereas the original definition of current flow betweenness centrality relies on the computation of random walks between all pairs of nodes in the network, the calculation of WFEBC requires that source and target nodes belong to restricted sets defined in accordance with the physical nature of the nodes in the real system.
- The unit supply between a given pair of source-target (s-t) nodes is weighted by the estimated demand of the target node as well as the relative capacity of the source node, compared to the total capacity of the network.

This allows WFEBC to go beyond traditional redundancy analysis and provide a measure of the capacity and criticality of a network's edges.

WFEBC was applied to the analysis of Net3, a case study network of 92 nodes, 5 sources or reservoirs and 119 links provided with the open-source hydraulic solver software EPANET 2.0 [4]. The measures of water flow edge betweenness centrality of the links (Figure 1) are compared to the results of a hydraulic resilience analysis performed on the network, using reserve capacity [7] as an index for link criticality. The darker the edge, the more critical.

Branches of tree structures connecting leaf nodes and bottlenecks of customer supply are excluded from the analysis: disconnections of single customers, regardless of their importance, causes the reserve capacity of the network to drop to 0, inconsistently granting highest criticality levels to links on the extremities. Moreover, due to their particular nature, the impact of the failure of such links is easy to comprehend from visual inspection of the structure of the network: assessment of their criticality is left to engineering judgment, based on the importance of the disconnected customers of amount of unmet demand.

The analysis of the reserve capacity and WFEBC values of the remaining links shows that:

- The hydraulic criticality analysis identifies 17 links of which a failure causes the reserve capacity of the network to drop significantly. Likewise, the graph theoretical analysis of Net3 correctly identifies the same 17 links as having the highest WFEBC values, with only slight variations in their ranking order.
- The remaining links show indiscernibly high (resp. low) values of reserve capacity (resp. WFEBC), their failure still allowing the network to meet up to 190% of the its nominal demand. Being almost all equivalently critical with regards to both metrics, their rankings show little correlation.

The results of the application of WFEBC to the analysis of Net3 are summarized by Figure 2, which compares the degrees of criticality of network links for reserve capacity and WFEBC.

To assess the added value of WFEBC compared to existing graph theoretic metrics, the shortest-path betweenness centrality values of the edges of the network are also calculated. The results show that, unlike WFEBC, shortest-path betweenness centrality fails to account for alternative paths and to identify some of the most critical links connecting the sources to the rest of the network as, by definition, it does not discriminate between source and target nodes.

The properties of WFEBC allow for two applications, depending on the level of knowledge the operator has of the network:

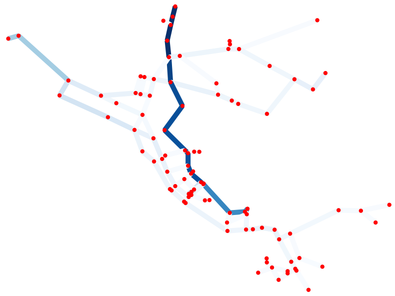


Fig. 1. WFEBC of Net3 links. The darker the link, the more central it is.

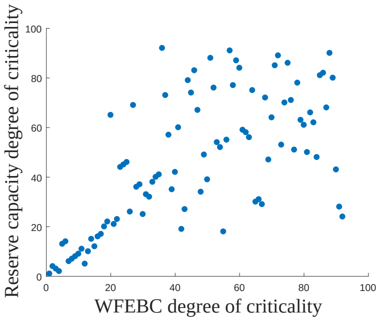


Fig. 2. Hydraulic criticality degree plotted against WFEBC criticality degree for Net3 links

- If the information required to run the hydraulic model and carry out a critical link analysis of the network is available, then computing an initial measure of the WFEBC provides a first level of understanding of the network's resilience and reduces the set of candidate links for further criticality analysis.
- On the other hand, WFEBC is also a self-standing metric that can be used to carry out resilience analyses when little is known about the system's dynamics or when the confidence in the hydraulic model is too low.

Further work could include integrating WFEBC results in various network optimisation or hydraulic model calibration problems.

References

1. Brandes, U., Fleischer, D.: Centrality measures based on current flow. Lecture Notes in Computer Science pp. 533–544 (2005), http://www.inf.uni-konstanz.de/algo/publications/bf-cmbcf-05.pdf
2. Goulter, I.C., Bouchart, F.: Reliability-constrained pipe network model. Journal of Hydraulic Engineering 116(2), 211–229 (1990)
3. Herrera, M., Abraham, E., Stoianov, I.: Graph-theoretic surrogate measures for analysing the resilience of water distribution networks. Procedia Engineering 119, 1241–1248 (2015), http://dx.doi.org/10.1016/j.proeng.2015.08.985
4. L. A. Rossman: Epanet 2 users manual (September) (2000)
5. Newman, M.E.J.: A measure of betweenness centrality based on random walks. Social networks 27(1), 39–54 (2005)
6. Todini, E.: Looped water distribution networks design using a resilience index based heuristic approach. Urban water 2(2), 115–122 (2000)
7. Wright, R., Herrera, M., Parpas, P., Stoianov, I.: Hydraulic resilience index for the critical link analysis of multi-feed water distribution networks. Procedia Engineering 119, 1249–1258 (2015), http://dx.doi.org/10.1016/j.proeng.2015.08.987

Part XII

Social and Political Networks

Multichannel Social Signatures and Persistent Features of Ego Networks

Sara Heydari[1], Sam G.B. Roberts[2], R.I.M. Dunbar[1,3], and Jari Saramäki[1]

[1] Department of Computer Science, Aalto University, Espoo, Finland,
sara.heydari@aalto.fi,
WWW home page: http://cs.aalto.fi/
[2] Department of Psychology, University of Chester, UK
[3] Department of Experimental Psychology, University of Oxford, UK

Social relationships that are strong and supportive are fundamentally important for health and well-being, in both humans and other primates [1, 2]. While close, emotionally intense relationships provide support and cohesion, weaker ties have been associated with beneficial diversity and access to resources outside one's everyday social circles. At the same time, maintaining social ties comes at a cost: time and cognitive resources are finite [3]. This cost is particularly high for close relationships [4]. Therefore personal networks typically have only a few close ties and many weak ties. This is visible both at the level of entire social networks [5] as well as in how individuals structure their personal networks[6].

In Ref. [6], it was shown that people allocate their mobile telephone calls to their alters rather inhomogeneously: a few closest alters get a disproportionate fraction of calls. Further, each individual was seen to have their own *social signature* that quantifies this inhomogeneity. Social signatures measure the fraction of communication targeted at alters of each rank, when the alters are ranked according to this fraction. In other words, they depict rank-frequency relationships of alters. Social signatures persist in time, even when there is heavy turnover in the ego network.

Besides calls, social relationships are shaped and maintained through many other communication channels. Since channels are different in their nature and functionality, people do not use them interchangeably. Several factors contribute to the choice of channel for each communication event, from the type of relationship to the aim of communication.

In order to understand the properties of ego networks and social signatures better, it is therefore important to look at data from multiple channels of communication. Comparing and combining information on different channels can be problematic because of their intrinsic differences. We suggest a method for constructing weighted ego networks from time-stamped communication data that makes different channels comparable (see Fig. 1), and allows constructing multichannel social signatures. We observe that the two social signatures that reflect call and SMS ego networks are very similar for each ego. At the same time, their composition in terms of alters may largely differ, which makes this ego-level similarity unexpected. Moreover, we observe that both single-channel and multi-channel signatures are persistent in time as observed earlier for calls-only signatures.

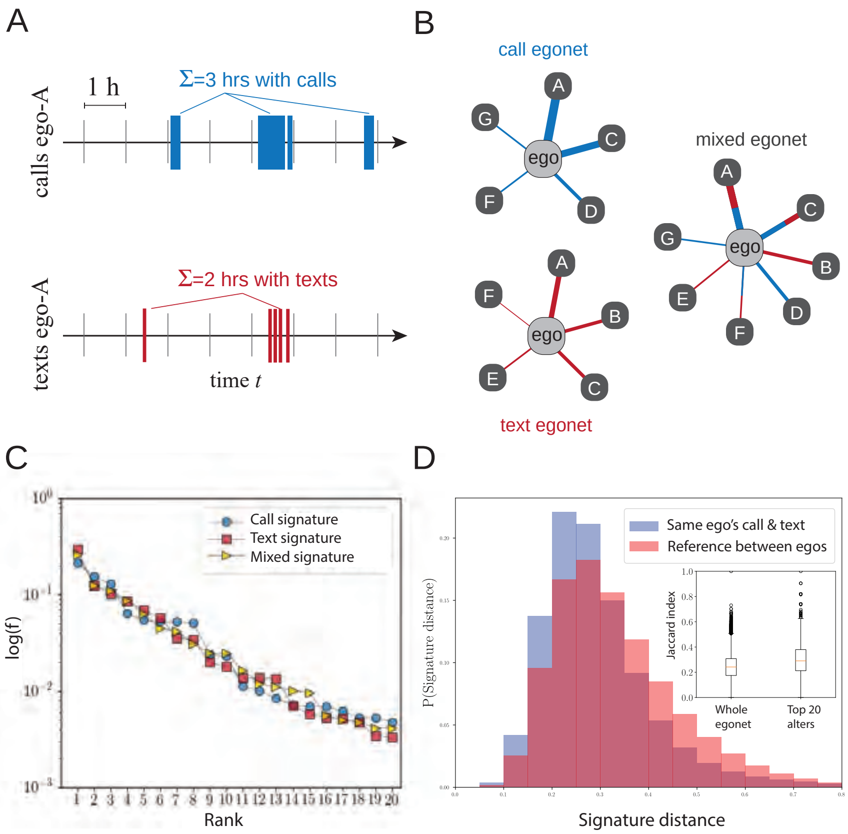


Fig. 1. Social signatures constructed from calls, texts, and both calls and texts show similar shapes. A) and B) The numbers of one-hour time bins that contain calls or texts are used as link weights in egocentric networks; these are more comparable than numbers of calls and texts whose ”units” do not match because one conversation may require large numbers of texts but just one call. C) Shapes of call, text, and mixed signatures for one individual (fraction of link weight as a function of alter rank). D) Distances between call and text signatures of individuals versus cross-individual distances show that each individual's call and text signatures tend to be similar. Inset: small values of Jaccard indices between sets of called and texted alters indicate differences in the composition of call/text ego networks.

Our results point towards the possibility of individuals having intrinsic ways of shaping their personal networks that are reflected on any of the communication channels they use; at the same time, the choice of channel for each of their alters appears independent of these ways.

References

1. Wittig, R.M., Crockford, C., Lehmann, J., Whitten, P.L., Seyfarth, R.M., Cheney, D.L.: Focused grooming networks and stress alleviation in wild female baboons, Hormones and Behavior 54, 170–177 (2008)
2. Holt-Lunstad, J., Smith, T.B., Layton, J.B.: Social relationships and mortality risk: A meta-analytic review, PLoS Medicine 7, e1000316 (2010)
3. Miritello, G., Moro, E., Lara, R., Martínez-López, R., Belchamber, J., Roberts, S.G.B., Dunbar, R.I.M: Time as a limited resource: Communication strategy in mobile phone networks, Social Networks 35, 89–95 (2013)
4. Roberts, S.G.B., Dunbar, R.I.M, Pollet, T.V., Kuppens, T.: Exploring variation in active network size: Constraints and ego characteristics, Social Networks 31, 138–146 (2009)
5. Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.-L.: Structure and tie strengths in mobile communication networks, Proceedings of the National Academy of Sciences (USA) 104, 7332–7336 (2007)
6. Saramäki, J., Leicht, E.A., López, E., Roberts, S.G.B, Reed-Tsochas, F., Dunbar, R.I.M.: Persistence of social signatures in human communication, Proceedings of the National Academy of Sciences 111, 942–947 (2014)

Associative nature of event-driven social dynamics: a network theory approach

Marija Mitrović Dankulov[1] and Jelena Smiljanić[1]

Scientific Computing Laboratory,
Center for the Study of Complex Systems, Institute of Physics Belgrade, University of Belgrade,
Pregrevica 118, 11080 Belgrade, Serbia
mitrovic@ipb.ac.rs
jelena.smiljanic@ipb.ac.rs

The emergence of collective social behavior in various social groups has attracted a lot of attention of researchers from the field commonly known as computational social science [2, 1, 5]. They combine the techniques from different areas of science, including statistical physics, complex network theory, and computer science, to quantitatively describe the dynamics and structure of various social groups, and to discover the underlying mechanisms. The abundance of data about human online behavior has enabled extensive studies of human activity patterns, social networks structure, as well as the emergence of collective behavior in online social groups. On the other hand, the growth and evolution of offline social communities, especially those with event-driven dynamics, have attracted a relatively little attention, mostly due to the lack of data. Many offline social groups have event-driven dynamics, i.e., their members meet and build social connections during the events which are well localized in time and space. These groups have an important role in every society since they include all spheres of social community life, for instance, social support groups, political campaigns and movements, leisure groups such as book clubs [7], or professional groups such as conferences [6]. Although these groups are inherently different considering their topic, type of activity or profile of their members, they all have event-driven dynamics which is responsible for the universal patterns of member's participations in group activities [6, 7].

Here we demonstrate this universality by analyzing the data from two different types of social groups: series of scientific conferences [6], which are representatives of event-driven professional social groups, and four leisure groups from Meetup platform [7]. We collected and curated the data for six different series of conferences from various fields of science [6]: American Physical Society March Meeting (APSMM), American Physical Society April Meeting, Society for Industrial and Applied Mathematics Annual Meetings, Neural Information Processing Systems Conference, International Conference on Supercomputing, and Annual International Conference on Research in Computational Molecular Biology. The data for four large Meetup groups, each of them belonging to a different category and having a different type of activity, have been collected using Meetup API [7]: Geamclt group is made of foodie thrill-seekers, VegasHiker (LVHK) group consists of hikers, Pittsburgh-free people search for free social events, and TechLife Columbus a technology-related community. For both social group types, we collected the list of their members, and events (conferences or meetups depending on the type of social group), as well as the attendance list for each event.

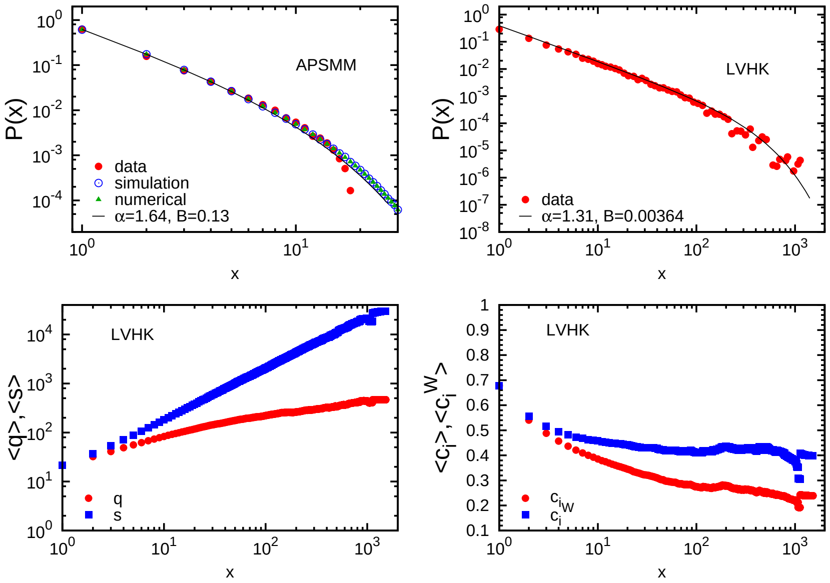


Fig. 1. (top) Probability distributions $P(x)$ of total number of participations x, for APSMM (left) and LVHK (right). Blue circles and green triangles in the left panel correspond to simulations and numerical solution of non-linear Polya urn model for APSMM. Solid line represents best fit to truncated power law distribution $x^{-\alpha}e^{-\beta x}$. (bottom) Dependence of members average degree $\langle q\rangle$ and strength $\langle s\rangle$ (left), and non-weighted $\langle c_i\rangle$ and weighted clustering coefficients $\langle c_i^W\rangle$ (right) on number of attended group events x for group LVHK.

This has allowed us to analyze in detail the participation patterns of members of these groups. Specifically, we have calculated the distributions of the total number of participations, the number and the time lag between two successive participations. All these distributions exhibit truncated power-law behaviour with the value of power-law exponent between 1 and 2, see Figure 1 (top). We model these event-driven dynamics using non-linear Polya urn model and show that the probability of member to attend the next event depends on the balance between the number of previously of attended and non-attended events through positive feedback mechanism. This suggests that event-driven dynamics is strongly influenced by social factors, such as members association with the community and inclusiveness of social groups.

To further explore this hypothesis, we analyze the evolution of ego-social networks of members of four Meetup groups. We map the data to a bipartite network of members and events, where the link between nodes i and j indicates the participation of member i in the event j. The social network between members of one Meetup group is obtained by projecting the appropriate bipartite network to members partition and filtering out the redundant links using the technique based on configuration model of random bipartite networks [4, 3]. Then, we study the evolution of average local features of ego-networks, such as degree, strength, weighted and non-weighted clustering coefficient, with the number of attended events, see Figure 1 (bottom). Our results show members increasing engagement in the group activities is primarily associated with the strengthening of already existing ties and increase in the bonding social capital.

References

1. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Rev. Mod. Phys. 81, 591–646 (May 2009)
2. Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al.: Life in the network: the coming age of computational social science. Science 323(5915), 721 (2009)
3. Saracco, F., Di Clemente, R., Gabrielli, A., Squartini, T.: Randomizing bipartite networks: the case of the world trade web. Scientific Reports 5, 10595 (2015)
4. Saracco, F., Straka, M.J., Di Clemente, R., Gabrielli, A., Caldarelli, G., Squartini, T.: Inferring monopartite projections of bipartite networks: an entropy-based approach. New Journal of Physics 19(5), 053022 (2017)
5. Sen, P., Chakrabarti, B.K.: Sociophysics: an introduction. Oxford University Press (2013)
6. Smiljanić, J., Chatterjee, A., Kauppinen, T., Mitrović Dankulov, M.: A theoretical model for the associative nature of conference participation. PLoS ONE 11(2), 1–12 (02 2016)
7. Smiljanić, J., Dankulov, M.M.: Associative nature of event participation dynamics: A network theory approach. PloS one 12(2), e0171565 (2017)

Predicting Offline Political Support with Online Behavioral Traces

Giona Casiraghi, Simon Schweighofer, Frank Schweitzer

ETH Zürich, Chair of System Design, Weinbergstrasse 56/58, 8092 Zürich, Switzerland,
giona@ethz.ch

1 Introduction

The internet age has brought with it a host of data, opening up vast new research possibilities for political scientists, and social scientists in general. However, these new possibilities also entail some methodological problems. First and foremost, there is the issue of *external validity* - in how far does our online data allow us to make any conclusions about the 'real world'? All too often online traces are naïvely equated with offline behavior. In political science, it is of course mostly the offline behavior - voting, lawmaking, governing - which we are really interested in.

2 Data and Methods

In this study, we overlay two datasets, both containing *behavioral data* of Swiss politicians. The first dataset – `Politnetz` – captures their *online behavior* on a Swiss political debate and networking platform. On `Politnetz`, politicians, as well as normal citizens can take a stand on various policy issues and debate them with each other. Crucially, they can also establish *support-relations* with each other, similar to the establishment of 'friendship' links on Facebook.

The second dataset captures politicians' *offline behavior* directly at their workplace, the Swiss parliament. More precisely, it contains their co-sponsorship relations, with which a parliamentary delegate can express support for the legislative proposal of other delegates. It has been shown previously that co-sponsorships are relevant indicators of political alliances, and can predict parliamentary voting behavior [3, 6, 4, 5].

Our offline dataset can be represented by a directed multi-edge network, where nodes are politicians and edges represent co-sponsorships links. It is multi-edge, because a delegate i can co-sponsor more than one legislative proposal of a delegate j during a legislation. In contrast, our online dataset is a directed, dichotomous network of support-links, meaning that these links can only be made once. Crucially, we can match politicians present in both datasets constructing a *multiplex network*, where one layer is the co-sponsorship network and the other layer is built from the support-relations expressed on the online platform.

In this paper we address the following question: can we predict the extent that a politician i co-sponsors a politician j in parliament by the information contained in the `Politnetz` support links? In other words: does online support tell us anything about 'real world' support?

To conduct analyses like this, where the dependent variable is not an individual property, but a multi-edge network, our chair recently developed a method for *multiplex network regression*, based on generalized hypergeometrical ensembles [2, 1]. Analogously to simple linear regression, this method can determine the influence of one or several independent variables on a multi-edge network, where such independent variables can be other multi-edge, dichotomous, or weighted networks. Like linear regression, it returns the significance values for the parameters, and allows for the computation of a pseudo-R^2 for the estimation of model quality.

Different from standard linear regression, it takes into account the intrinsic interdependency of observations typical of networks and relational datasets. It accomplishes this by estimating the amount of edges between nodes i and j that we can expect at random, given the in- and out-degree of both i and j. It then estimates in how far the independent variable(s) can explain deviations from this random estimate.

In a first step, we will use our network regression method to regress the dichotomous network of `Politnetz` support links on the multi-edge co-sponsorship network from the Swiss parliament. Our hypothesis is: If politician i established a support relation to politician j on `Politnetz`, there will be more co-sponsorship relations from i to j in the parliament than expected at random.

In a second step, we will introduce a somewhat more sophisticated model. Instead of simply using `Politnetz` support links as independent variables, we will use them to determine similarity between politicians. Two politicians shall be maximally similar if they give their support to the same people on `Politnetz`, and maximally dissimilar if they give it to completely different people. For this purpose, we use Jaccard similarity. The similarity between two politicians i and j is defined as follows, where S_i (S_j) refers to the set of politicians supported by i (j respectively):

$$J(i,j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} = \frac{|S_i \cap S_j|}{|S_i| + |S_j| - |S_i \cap S_j|} \tag{1}$$

The matrix of similarities between all pairs of politicians is then used as independent variable, with the hypothesis being: If two politicians i and j are more similar with respect to whom they are supported from on `Politnetz`, there will be more co-sponsorship relations between i and j in the parliament than expected at random.

3 Outcomes

The first network regression model, using `Politnetz` support links as independent variable, retrieved a significant effect ($\beta = .6, p < .001$). However, the pseudo-R^2 is only .003. Thus, we have to conclude that, even though there is a statistically significant positive effect of `Politnetz` support links on parliamentary co-sponsorships, the explanatory value of the model is vanishingly small.

The second model, using `Politnetz` support based Jaccard similarity as independent variable, also retrieved a statistically significant, positive effect ($\beta = .3, p < .001$). This time, however, the new model also has considerable explanatory power, with a pseudo-R^2 of .152.

4 Conclusion

The question ’can we predict offline political support with online behavioral traces’ cannot be answered with a simple yes or no. What our study shows is that naîve assumptions, such as ’online support equals offline support’, may fail. While online support-links do have an influence on co-sponsorships in the Swiss parliament, they do not *explain* co-sponsorships. This, however, does not mean that online behavioral traces have no explanatory power: If we use `Politnetz` support links to determine political similarity between politicians, and if we further assume that political similarity leads to more parliamentary co-sponsorships, we can construct a model with much better explanatory power. We are looking forward to extracting more independent variables from the `Politnetz` dataset and integrating them into our model, as well as to applying our method to different contexts.

References

1. Giona Casiraghi. Multiplex Network Regression: How do relations drive interactions? *arXiv preprint arXiv:1702.02048*, feb 2017.
2. Giona Casiraghi, Vahan Nanumyan, Ingo Scholtes, and Frank Schweitzer. Generalized Hypergeometric Ensembles: Statistical Hypothesis Testing in Complex Networks. *arXiv preprint arXiv:1607.02441*, jul 2016.
3. James H Fowler. Connecting the Congress: A study of cosponsorship networks. *Political Analysis*, 14(4):456–487, 2006.
4. Daniel Kessler and Keith Krehbiel. Dynamics of cosponsorship. *American Political Science Review*, 90(03):555–566, 1996.
5. Michele L Swers. Transforming the agenda. In *Women Transforming Congress*, volume 4, page 260. University of Oklahoma Press, 2002.
6. Jeffery C Talbert and Matthew Potoski. Setting the legislative agenda: The dimensional structure of bill cosponsoring and floor voting. *The Journal of Politics*, 64(03):864–891, 2002.

Organization & Committees

General Chair	Hocine Cherifi, University of Burgundy, France
Advisory Board	Raissa D'Souza, University of California, Davis, USA Sabrina Gaito, University of Milan, Italy Ben Y. Zhao, University of Chicago, USA
Program Co-Chairs	Chantal Cherifi, University of Lyon 2, France Mirco Musolesi, University College London, UK Márton Karsai, ENS de Lyon, France
Poster Chairs	Hamamache Kheddouci, University of Lyon 1, France Huijuan Wang, Delft University of Technology, Netherlands
Publicity Chair	Bruno Gonçalves, New York University, USA Feng Xia, Dalian University of Technology, China Carlo Piccardi, Politecnico di Milano
Local Committee	Lounes Bentaha, University of Lyon 2, France Chantal Cherifi, University of Lyon 2, France Jannik Laval ,University of Lyon 2, France
Publication Chair	Sabrina Gaito, University of Milan, Italy
Tutorial Chair	Jinhu Lü, Chinese Academy of Sciences, China
Sponsor Chair	Eric Fleury, ENS de Lyon, France
Submission Chair	Christian Quadri, University of Milan, Italy
Web Chair	Matteo Zignani, University of Milan, Italy

Program Committee

Sophie Achard	GIPSA-Lab - CNRS, France
Masaki Aida	Tokio Metropolitan University, Japan
Luca Maria Aiello	Nokia Bell Labs, UK
Tatsuya Akutsu	Kyoto University, Japan
Reka Albert	Pennsylvania State University, USA
Antoine Allard	Centre de Recerca Matemàtica, Spain
Eivind Almaas	Norwegian University of Science and Technology, Norway
Claudio Altafini	Linköping University, Sweden
Lucila Alvarez-Zuzek	Universidad Nacional de Mar del Plata, Argentina
Fred Amblard	University Toulouse 1 Capitole, France
Claudio Angione	Teesside University, UK
Alberto Antonioni	Carlos III University of Madrid, Spain
Nino Antulov-Fantulin	ETH Zurich, Switzerland
Nuno Araujo	Universidade de Lisboa, Portugal
Elsa Arcaute	University College London, UK
Valerio Arnaboldi	IIT-CNR, Italy
Tomaso Aste	University College London, UK
Martin Atzmueller	Tilburg University, Netherlands
Rodolfo Baggio	Bocconi University, Italy
James Bagrow	University of Vermont, USA
Sven Banisch	Max Planck Institute for Mathematics in the Sciences, Germany
Yaneer Bar-Yam	New England Complex Systems Institute, USA
Baruch Barzel	Bar-Ilan University, Israel
Nikita Basov	St. Petersburg State University, Russia
Gareth Baxter	University of Aveiro, Portugal
Mariano Beguerisse Diaz	University of Oxford, UK
Rosa M. Benito	Universidad Politecnica de Madrid (UPM), Spain
Jacob Biamonte	University of Malta, Malta
Ginestra Bianconi	Queen Mary University of London, UK
Jeremy Blackburn	University of Alabama at Birmingham, USA
Anthony Bonato	Ryerson University, Canada
Pierre Borgnat	CNRS, Laboratoire de Physique ENS de Lyon, France
Stefan Bornholdt	University of Bremen, Germany
Dan Braha	NECSI, USA
Ulrik Brandes	University of Konstanz, Germany
Lidia A. Braunstein	Universidad Nacional de Mar del Plata, Argentina
Markus Brede	University of Southampton, UK

Marco Bressan	Sapienza University of Rome, Italy
Dirk Brockmann	Humboldt University, Germany
Piotr Bródka	Wrocław University of Science and Technology, Poland
Javier M. Buldu	Universidad Rey Juan Carlos and Center for Biomedical Technology, Spain
Raffaella Burioni	Università di Parma, Italy
Kanat Camlibel	University of Groningen, Netherlands
Carlo Vittorio Cannistraci	Technical University Dresden, Germany
Vincenza Carchiolo	Universita di Catania, Italy
Alessio Cardillo	Catalan Institute of Human Paleoecology and Social Evolution (IPHES), Spain
Rui Carvalho	Durham University, UK
Giona Casiraghi	ETH Zurich, Switzerland
Remy Cazabet	UPMC - Paris, France
L. Elisa Celis	École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Mario Chavez	Lena - CNRS, France
Kwang-Cheng Chen	University of South Florida, USA
Fu Lai Chung	Hong Kong Polytechnic University, Hong Kong
Richard Clegg	Queen Mary University of London, UK
Jack Cole	ARL (ret), USA
Giacomo Como	Lund University, Sweden
Luciano da F. Costa	University of Sao Paulo, Brazil
Emanuele Cozzo	University of Zaragoza, Spain
Regino Criado	Universidad Rey Juan Carlos, Spain
Mihai Cucuringu	University of Oxford and Alan Turing Institute, UK
Jörn Davidsen	University of Calgary, Canada
Bhaskar Dasgupta	University of Illinois at Chicago, USA
Fabrizio De Vico Fallani	Inria - ICM, France
Michela Del Vicario	IMT Institute for Advanced Studies, Italy
Jean-Charles Delvenne	University of Louvain, Belgium
José Devezas	INESC TEC and DEI-FEUP, Portugal
Jana Diesner	University of Illinois at Urbana-Champaign, USA
Louis J. Dubé	Université Laval, Canada
Jordi Duch	Universitat Rovira i Virgili, Spain
Marten During	Université de Luxembourg, Luxembourg
Mohammed El Hassouni	FSR-UMV, Marocco
Omodei Elisa	UNICEF, USA
Frank Emmert-Streib	Tampere University of Technology, Finland
Gunes Ercal	SIUE, USA
Ernesto Estrada	University of Strathclyde, UK
Tim Evans	Imperial College London, UK

Mauro Faccin	ICTEAM, Universitè Catholique de Louvain, Belgium
Giorgio Fagiolo	Sant'Anna School of Advanced Studies, Italy
Hocine Ficheri	ENS, France
Alessandro Flammini	Indiana University, USA
Eric Fleury	ENS Lyon / INRIA, France
Manuel Foerster	University of Hamburg, Germany
Mattia Frasca	University of Catania, Italy
Sabrina Gaito	University of Milan, Italy
José Manuel Galán	University of Burgos, Spain
Edoardo Gallo	University of Cambridge, UK
Yérali Gandica	Université de Namur, Belgium
Antonios Garas	ETH Zurich, Switzerland
Alvaro Garcia-Recuero	Queen Mary University of London, UK
Gourab Ghoshal	University of Rochester, USA
Silvia Giordano	SUPSI, Switzerland
James Gleeson	University of Limerick, Ireland
Kwang-Il Goh	Korea University, Korea
Sergio Gómez	Universitat Rovira i Virgili, Spain
Jesus Gomez-Gardenes	University of Zaragoza, Spain
Bruno Gonçalves	New York University, USA
Przemyslaw Grabowicz	Max Planck Institute for Software Systems, Germany
Steve Gregory	University of Bristol, UK
Thilo Gross	University of Bristol, UK
Jelena Grujic	Vrije Universiteit Brussel, Belgium
Jean-Loup Guillaume	L3i - Université de la Rochelle, France
Mehmet Gunes	University of Nevada, Reno, USA
Aric Hagberg	Los Alamos National Laboratory, USA
Edwin Hancock	University of York, UK
Chris Hankin	Imperial College London, UK
Jin-Kao Hao	University of Angers, France
Yukio Hayashi	Japan Advanced Institute of Science and Technology, Japan
Laurent Hébert-Dufresne	Santa Fe Institute, USA
Denis Helic	KTI, TU-Graz, Austria
Shaun Hendy	University of Auckland, New Zealand
Babak Heydari	Stevens Institute of Technology, USA
Desmond Higham	University of Strathclyde, UK
Philipp Hoevel	TU Berlin, Germany
Seok-Hee Hong	University of Sydney, Australia
Ulrich Hoppe	University Duisburg-Essen, Germany
Pan Hui	Hong Kong University of Science and Technology, Hong Kong
Yuichi Ikeda	Kyoto University, Japan

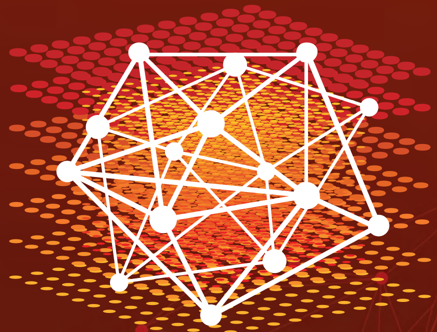
Gerardo Iñiguez	National Autonomous University of Mexico (UNAM), Mexico
Plamen Ivanov	Boston University and Harvard Medical School, USA
Mahdi Jalili	RMIT University, Australia
Marco Alberto Javarone	School of Computer Science, University of Hertfordshire, UK
Hawoong Jeong	Korea Advanced Institute of Science and Technology, Korea
Tao Jia	Southwest University, Chongqing, China
Di Jin	Tianjin University, China
Hang-Hyun Jo	Asia Pacific Center for Theoretical Physics, Korea
Nick Jones	Imperial College, USA
Bertrand Jouve	CNRS, France
Byungnam Kahng	Seoul National University, Korea
Rushed Kanawati	Université Paris 13, France
Márton Karsai	ENS de Lyon, France
Mehmet Kaya	Firat University, Turkey
Przemyslaw Kazienko	Wroclaw University of Technology, Poland
Dror Kenett	Boston University, USA
Yoed Kenett	University of Pennsylvania, USA
Khaldoun Khashanah	Stevens Institute of Technology, USA
Hamamache Kheddouci	Universit Claude Bernard, France
Hyoungshick Kim	Sungkyunkwan University, Korea
Maksim Kitsak	Northeastern University, UK
Mikko Kivela	University of Oxford, UK
Peter Klimek	Medical University of Vienna, Austria
Nicolas Kourtellis	Telefonica Research, Spain
Danai Koutra	University of Michigan, USA
Jérôme Kunegis	University of Namur, Belgium
Valentina Kuskova	National Research University, Russia
Ryszard Kutner	University of Warsaw, Poland
Haewoon Kwak	Qatar Computing Research Institute, Qatar
Lucas Lacasa	Queen Mary University of London, UK
Renaud Lambiotte	University of Oxford, UK
Christine Largeron	Université de Lyon, France
Jennifer Larson	New York University, USA
Matthieu Latapy	CNRS, France
Stephen Law	University College London, USA
Anna T. Lawniczak	University of Guelph, Canada
Benedicte Le-Grand	Université Paris 1 Panthéon Sorbonne, France
Eric Leclercq	University of Burgundy, France
Sang Hoon Lee	Korea Institute for Advanced Study, Korea

Benedicte Le-Grand	Université Paris 1 Panthépn Sorbonne, France
Sune Lehmann	Technical University of Denmark, Denmark
Xiang Li	Fudan University, China
Nelly Litvak	University of Twente, Netherlands
Yang-Yu Liu	Harvard Medical School, USA
Alessandro Longheu	University of Catania, Italy
Jinhu Lu	Chinese Academy of Sciences, China
John C.S. Lui	The Chinese University of Hong Kong, Hong Kong
Matteo Magnani	Uppsala University, Sweden
Clemence Magnien	CNRS - UPMC Sorbonne Universités, France
Hernan Makse	City College of New York, USA
Giuseppe Mangioni	University of Catania, Italy
Madhav Marathe	Virginia Tech, USA
Andrea Marino	University of Pisa, Italy
Antonio Marques	King Juan Carlos University, Spain
Michael Mäs	University of Groningen, Netherlands
Cristina Masoller	Universitat Politecnica de Catalunya, Spain
Rossana Mastrandrea	IMT Institute of Advanced Studies, Italy
Naoki Masuda	University of Bristol, UK
Petr Matous	University of Sydney, Australia
Matúš Medo	UESTC Chengdu, China
Natarajan Meghanathan	Jackson State University, USA
Guy Melançon	Université de Bordeaux, France
Jörg Menche	Austrian Academy of Sciences, Austria
Jose Fernando Mendes	University of Aveiro, Portugal
Ronaldo Menezes	Florida Institute of Technology, USA
Radosław Michalski	Wrocław University of Science and Technology, Poland
Bivas Mitra	Indian Institute of Technology Kharagpur, India
Marija Mitrovic	Institute of Physics Belgrade, Serbia
Suzy Moat	University of Warwick, UK
Yamir Moreno	Universidad de Zaragoza, Spain
Sotiris Moschoyiannis	University of Surrey, UK
Igor Mozetič	Jozef Stefan Institute, Slovenia
Animesh Mukherjee	Indian Institute of Technology, Kharagpur, India
Tsuyoshi Murata	Tokyo Institute of Technology, Japan
Katarzyna Musial	Bournemouth University, UK
Muaz Niazi	COMSTATS Institute of IT, Pakistan
Andrea Omicini	Università di Bologna, Italy
Gergely Palla	Eötvös University, Hungary

Pietro Panzarasa	Queen Mary University of London, UK
Fragkiskos Papadopoulos	Cyprus University of Technology, Cyprus
Symeon Papadopoulos	Information Technologies Institute, Greece
Michela Papandrea	SUPSI, Switzerland
Han Woo Park	YeungNam University, Korea
Juyong Park	Korea Advanced Institute of Science and Technology, Korea
Andrea Passarella	IIT-CNR, Italy
Leto Peel	Université Catholique de Louvain, Belgium
Tiago Peixoto	University of Bath, UK
Matjaz Perc	University of Maribor, Slovenia
Nicola Perra	University of Greenwich, UK
Giovanni Petri	ISI Foundation, Italy
Carlo Piccardi	Politecnico di Milano, Italy
Carlos Pineda	Universidad Nacional Autonoma de Mexico, Mexico
Sebastian Poledna	International Institute for Applied System Analysis, Austria
Chiara Poletto	INSERM UMR-S 1136, France
Victor Preciado	University of Pennsylvania, USA
Natasa Przulj	University College London, UK
Christian Quadri	University of Milan, Italy
Marco Quaggiotto	ISI Foundation, Italy
Walter Quattrociocchi	Labss,Institute of Cognitive Sciences and Technologies, Italy
Jose J. Ramasco	IFISC (CSIC-UIB), Spain
Asha Rao	RMIT University, Australia
Felix Redd-Tsochas	University of Oxford, UK
Gesine Reinert	University of Oxford, UK
Pedro Ribeiro	Universidade do Porto, Portugal
Massimo Riccaboni	IMT Institute for Advanced Studies, Italy
Laura Ricci	University of Pisa, Italy
Luis E C Rocha	Karolinska Institutet, Sweden
Luis M. Rocha	Indiana University, USA
Francisco Rodrigues	University of São Paulo, Brazil
Henrik Ronellenfitsch	Massachusetts Institute of Technology, USA
Luca Rossi	University of Copenhagen, Denmark
Martin Rosvall	Umeå Univeristy, Sweden
Camille Roth	CNRS, France
Amir Rubin	Ben-Gurion University of the Negev, Israël
Marc Santolini	Northeastern University, USA
Francisco C. Santos	INESC-ID and Instituto Superior Técnico, and ATP group, Portugal

Jari Saramäki	Aalto University, Finland
Hiroki Sayama	Binghamton University, SUNY, USA
Antonio Scala	Institute for Complex Systems / Italian National Research Council, Italy
Maximilian Schich	The University of Texas at Dallas, USA
Grant Schoenebeck	University of Michigan, USA
Ingo Scholtes	ETH Zurich, Switzerland
Frank Schweitzer	ETH Zurich, Switzerland
Caterina Scoglio	Kansas State University, USA
Simone Severini	University College London, UK
Aneesh Sharma	Twitter Inc, USA
Rajesh Sharma	University of Tartu, Estonia
Tiago Simas	Univerity of Cambridge, UK
Filippo Simini	University of Bristol, UK
Anurag Singh	National Institute of Technology Delhi, India
Per Sebastian Skardal	Trinity College, USA
Michael Small	The University of Western Australia, Australia
Zbigniew Smoreda	Orange Labs, France
Chaoming Song	University of Miami, USA
Mauro Sozio	Télécom ParisTech, France
Jie Sun	Clarkson University, USA
Pål Sundsøy	Norges Bank Investment Management, Norway
Michael Szell	Hungarian Academy of Sciences, Hungary
Bosiljka Tadic	Jozef Stefan Institute, Slovenia
Lucia Tajoli	Politecnico di Milano, Italy
Kazuhiro Takemoto	Kyushu Institute of Technology, Japan
Frank Takes	Leiden University, Netherlands
Fabien Tarissan	CNRS - ENS Paris-Saclay, France
Dane Taylor	University of Buffalo, USA
Claudio Juan Tessone	University of Zurich, Switzerland
I-Hsien Ting	National University of Kaohsiung, Taiwan
Olivier Togni	Burgundy University, France
Ljiljana Trajkovic	Simon Fraser University, Canada
Jan Treur	Vrije Universiteit Amsterdam, Netherlands
Milena Tsvetkova	London School of Economics and Political Science, UK
Liubov Tupikina	Ecole Polytechnique, France
Stephen Uzzo	New York Hall of Science, USA
Sergi Valverde	University Pompeu Fabra, Spain
Piet Van Mieghem	Delft University of Technology, Netherlands
Balazs Vedres	Central European University, Hungary

Huijuan Wang	Delft University of Technology, Netherlands
Pinghui Wang	Xi'an Jiaotong University, China
Xiaofan Wang	Shanghai Jiao Tong University, China
Guanghui Wen	Southeast University, Nanjing, China
Richard Wilson	University of York, UK
Ernst Wit	University of Groningen, Netherlands
Feng Xia	Dalian University of Technology, China
Gang Yan	Tongji University, China
Wenwu Yu	Southeast University, China
Maryam Zamani	Eötvös University, Ungheria
Ivan Zelinka	VSB Technical University of Ostrava, Czech Republic
Zi-Ke Zhang	Hangzhou Normal University, China
Matteo Zignani	University of Milan, Italy
Antonio Zippo	CNR, Italy
Vinko Zlatić	Rudjer Boskovic Institute, Croatia



COMPLEX NETWORKS 2017

ENS DE LYON

POLE R.T.I. RECHERCHE TRANSFERT INNOVATION

MaDICS

MILYON UNIVERSITE DE LYON

Fédération Informatique de Lyon

CNRS

UNIVERSITÀ DEGLI STUDI DI MILANO

La Région Auvergne-Rhône-Alpes

LIP

uB

disp

IXXI

UCB Lyon 1

UNIVERSITÉ DE LYON

LIRIS

Inria

GRANDLYON la métropole

TUDelft

Fst UCB Lyon 1

UCL

GDR RSD

L2i

UNIVERSITÉ LUMIÈRE LYON 2

TRIBE COMMUNICATION

SHAMAN

The 6th International Conference on Complex Networks and Their Applications
November 29 - December 01 - Lyon, France

www.complexnetworks.org

ISBN 978-2-9557050-2-5