Network analysis of (big) data

José Lages

Equipe PhAs / Theoretical Physics and Astrophysics group
Institut UTINAM / OSU THETA / CNRS / Université de Bourgogne-Franche-Comté

DataBFC - ouvrir et gérer les données de la recherche en Bourgogne-Franche-Comté,
November 14th 2017, Besançon

Collaborators

Laboratoire de Physique Théorique de Toulouse / UPS / CNRS

Dima Shepelyansky, Klaus Frahm

Institut Curie / PSL / Inserm

Andrei Zinovyev (Institut Curie / PSL / Inserm)

Institut UTINAM / UBFC /CNRS

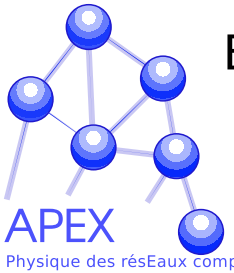
José Lages, Célestin Coquidé, Guillaume Rollin

Projects

CNRS MI MASTODONS

ApliGoogle

Applications of Google matrix to directed networks and Big Data

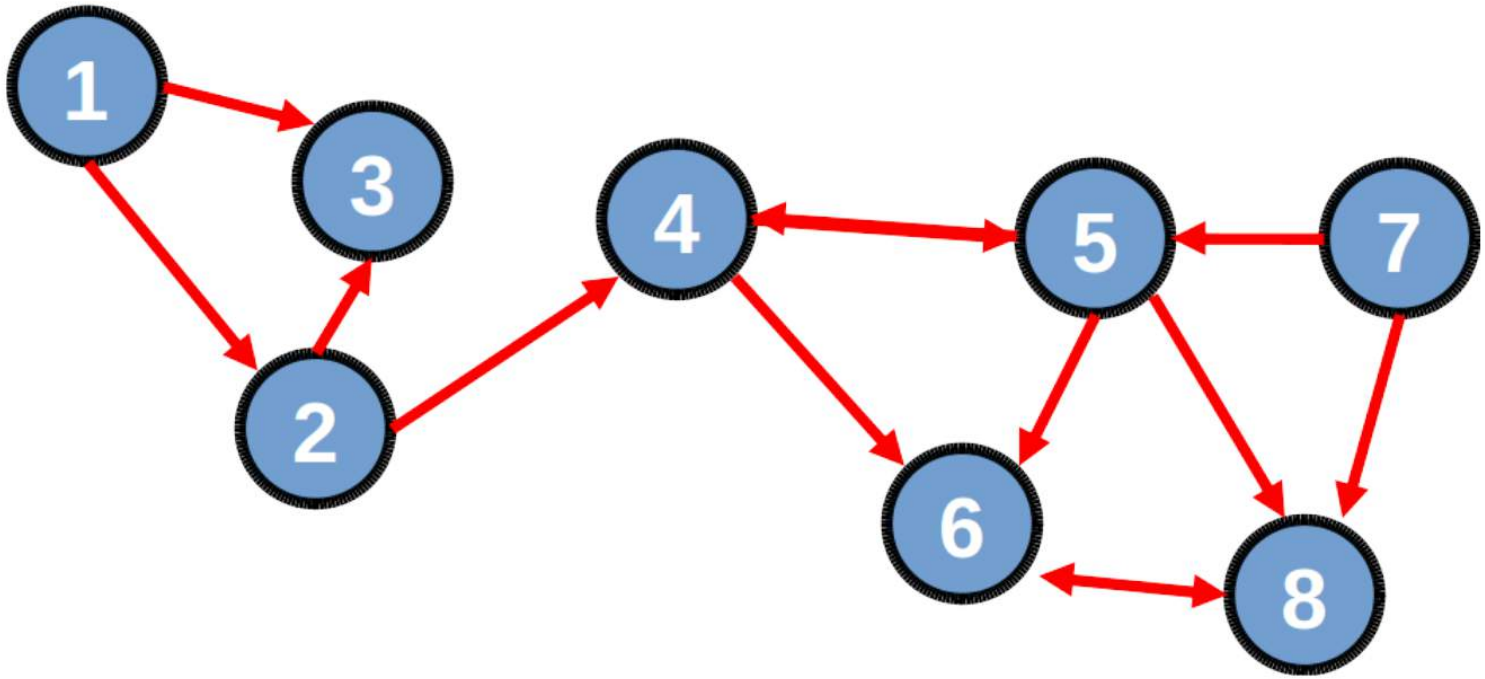


BFC Region project

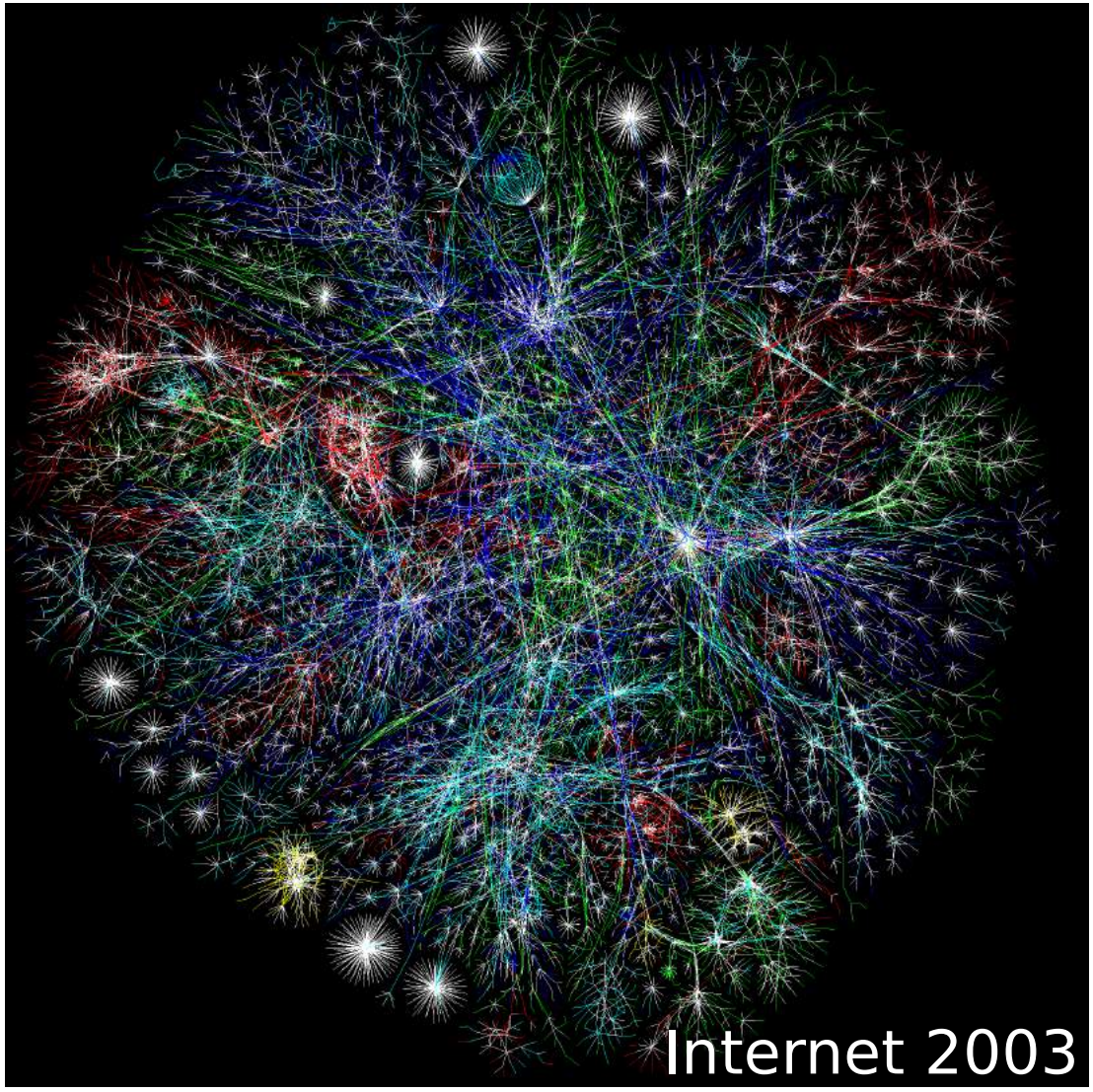
RÉGION BOURGOGNE FRANCHE COMTÉ

Network

nodes + (weighted) directed links



Complex network

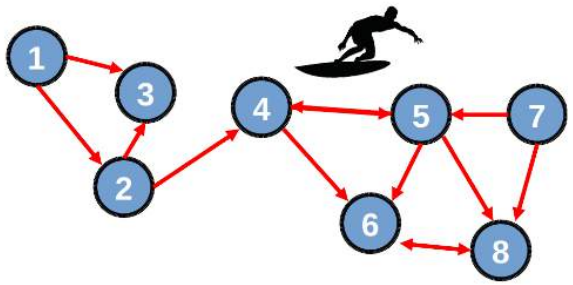


Big Data seen as directed networks

Non exhaustive list of applications

Data	Nodes	Links
WWW	Webpages	Hyperlinks
Wikipedia	Wikiarticles	Intrawiki citations
Social networks	Members	Acquaintances
World Trade	Goods x countries	Economic fluxes
Omics	Proteins	Causal relations / Interactions
Linux	Kernel commands	Command successions
DNA	Words of letters A,T,G,C	Word successions
Brain	Nerve cells	Axons
Go game	Plaquettes / patterns	Pattern successions
Cosmic web	Sub-halos	Proximity rules

What is the (nth) most important node ?



Adjacency matrix

$$A_{ij} = \begin{cases} 0 & \text{if } j \nrightarrow i \\ 1 & \text{if } j \rightarrow i \end{cases}$$

$$\mathbf{A} = \begin{pmatrix} 0&0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0 \\ 0&1&0&0&1&0&0&0 \\ 0&0&0&1&0&0&1&0 \\ 0&0&0&1&1&0&0&1 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&1&0 \end{pmatrix}$$

Stochastic matrix

$$S_{ij} = \begin{cases} A_{ij}/k_{out}(j) & \text{if } k_{out}(j) \neq 0 \\ 1/N & \text{otherwise} \end{cases}$$

node j outdegree $k_{out}(j) = \sum_{i=1}^{N} A_{ij}$

$$\mathbf{S} = \begin{pmatrix} 0&0&1/8&0&0&0&0&0 \\ 1/2&0&1/8&0&0&0&0&0 \\ 1/2&1/2&1/8&0&0&0&0&0 \\ 0&1/2&1/8&0&1/3&0&0&0 \\ 0&0&1/8&1/2&0&0&1/2&0 \\ 0&0&1/8&1/2&1/3&0&0&1 \\ 0&0&1/8&0&0&0&0&0 \\ 0&0&1/8&0&1/3&1&1/2&0 \end{pmatrix}$$

Google matrix

$$G_{ij} = \alpha S_{ij} + (1-\alpha)/N$$

Damping factor $\alpha = 0.85$

$\alpha = 0.8$

$$\mathbf{G} = \begin{pmatrix} 1/40&1/40&1/8&1/40&1/40&1/40&1/40&1/40 \\ 17/40&1/40&1/8&1/40&1/40&1/40&1/40&1/40 \\ 17/40&17/40&1/8&1/40&1/40&1/40&1/40&1/40 \\ 1/40&17/40&1/8&1/40&7/24&1/40&1/40&1/40 \\ 1/40&1/40&1/8&17/40&1/40&1/40&17/40&1/40 \\ 1/40&1/40&1/8&17/40&7/24&1/40&1/40&33/40 \\ 1/40&1/40&1/8&1/40&1/40&1/40&1/40&1/40 \\ 1/40&1/40&1/8&1/40&7/24&33/40&17/40&1/40 \end{pmatrix}$$

PageRank algorithm

More a given node is pointed by important nodes
more this node is important
(Measure of influence)

$$\mathbf{P}(t) = \underbrace{GG\ldots G}_{t\,\text{times}}\mathbf{P}(0) = G^t\mathbf{P}(0)$$

$P_i(t)$ is the probability that the random surfer ends at node i after t steps

Providing $\alpha < 1$, $\mathbf{P}(t)$ converges to a unique PageRank vector $\mathbf{P}$

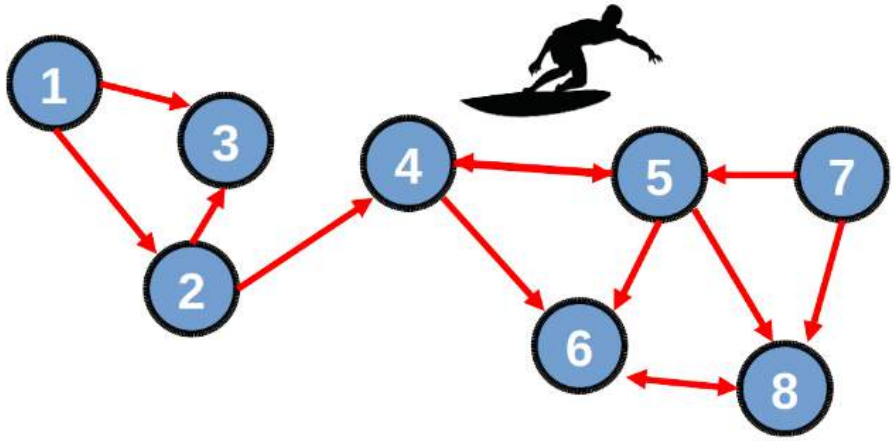
$$G\mathbf{P} = \mathbf{P}$$

After a sufficiently long journey, P_i is the probability that a random surfer ends at node i

The PageRank index is $K \in \{1, \ldots, N\}$
$K = 1$ for page with highest P_i
$K = N$ for page with lowest P_i

PageRank algorithm is at the heart of
Google
search engine
(Brin & Page '98)

What is the (nth) most important node ?



$$\mathbf{P} = \begin{pmatrix} 0.0310945256873059\mathbf{7} \\ 0.0435323361475661\mathbf{7} \\ 0.0609452708660655\mathbf{8} \\ 0.0672941236179782\mathbf{6} \\ 0.0704499859958617\mathbf{1} \\ \mathbf{0.351816793560944\mathbf{89}} \\ 0.0310945256873059\mathbf{7} \\ 0.3437724384369714\mathbf{3} \end{pmatrix}$$

Distribution $P(K)$
where K is the rank index:

$P(1) = \mathbf{0.351816793560944\mathbf{89}}$
$P(2) = 0.3437724384369714\mathbf{3}$
$P(3) = 0.0704499859958617\mathbf{1}$
$P(4) = 0.0672941236179782\mathbf{6}$
$P(5) = 0.0609452708660655\mathbf{8}$
$P(6) = 0.0435323361475661\mathbf{7}$
$P(7) = P(8) = 0.0310945256873059\mathbf{7}$

Rankings of World Universities

ACADEMIC RANKING OF WORLD UNIVERSITIES SINCE 2003

THE WORLD UNIVERSITY RANKINGS

QS WORLD UNIVERSITY RANKINGS®

⋮

About **20 different** global university rankings are listed in the Wikipedia page "College and university rankings"

All these rankings are composite:

α 

$+\beta$ 

$+\gamma$ 

$+\delta$ 

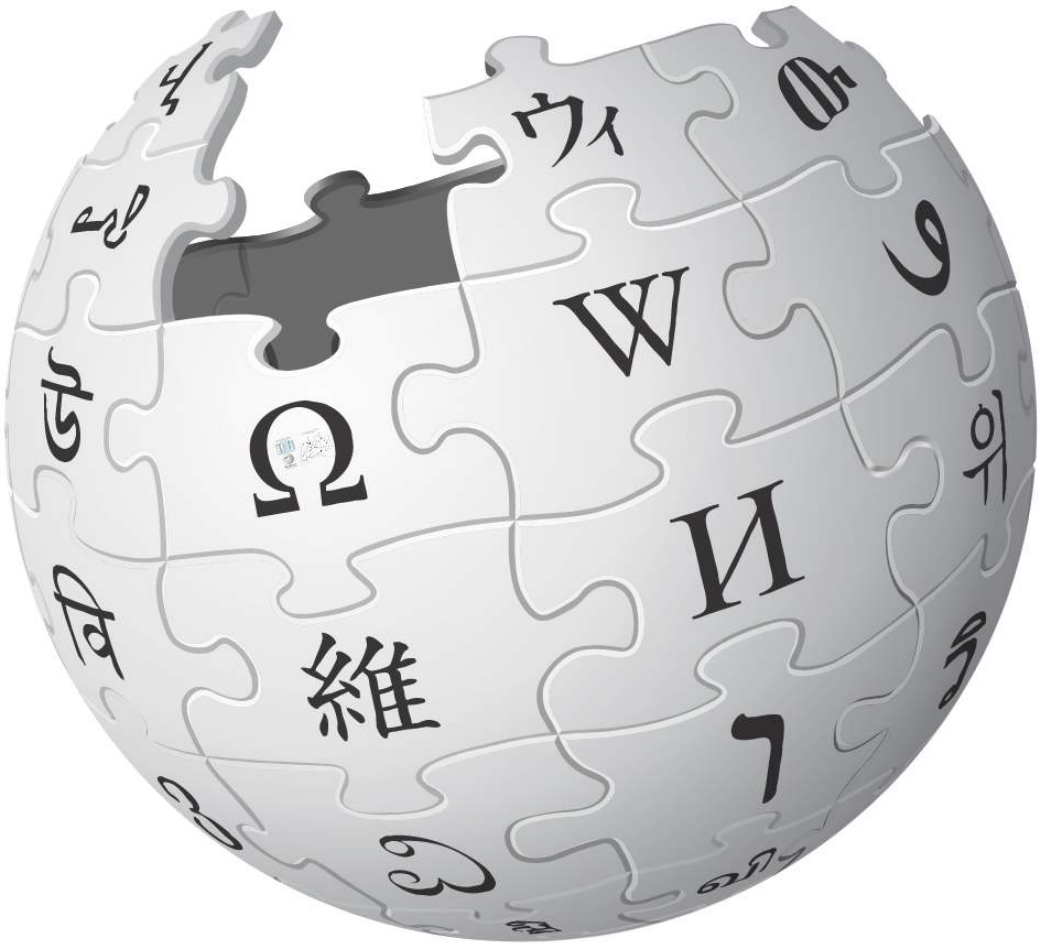
$+ \ldots$

Composite score 

Also, universities are preselected

These rankings have an impact on scientific and educational policies of governments

Is
there
an universal
ranking without
a priori criteria and
without cultural bias ?



WIKIPEDIA

The Free Encyclopedia

(Most of) human knowledge
is encoded in Wikipedia

Everybody use it
at least as a first approach
=
First contact with a subject

About 40M wikipages
280 language editions

24 Wikipedia language editions

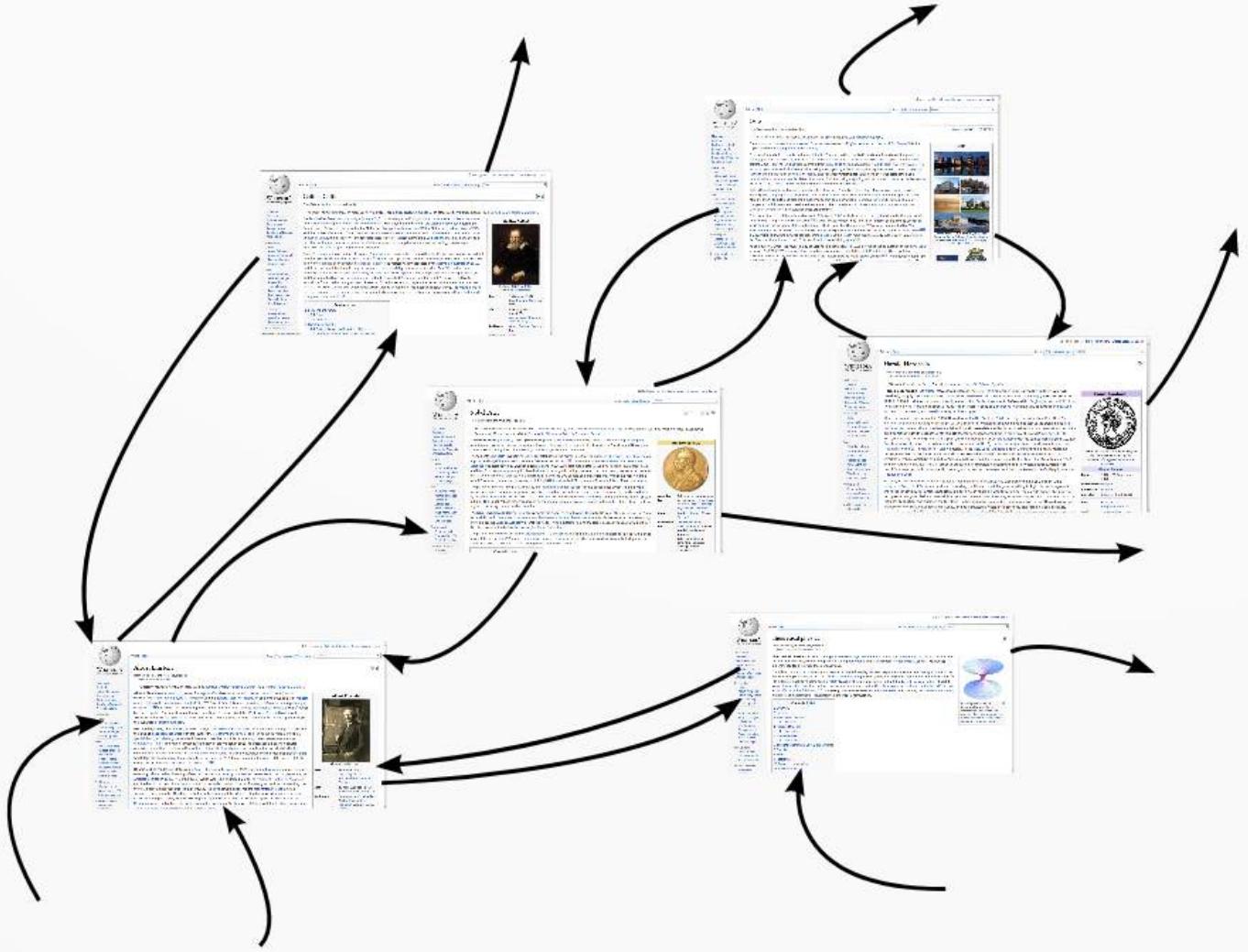
covering 59% of world population
and 68% total Wikipages

Edition	Language	N	Edition	Language	N
EN	English	4 212 493	VI	Vietnamese	594 089
DE	German	1 532 978	FA	Persian	295 696
FR	French	1 352 825	HU	Hungarian	235 212
NL	Dutch	1 144 615	KO	Korean	231 959
IT	Italian	1 017 953	TR	Turkish	206 311
ES	Spanish	974 025	AR	Arabic	203 328
RU	Russian	966 284	MS	Malaysian	180 886
PL	Polish	949 153	DA	Danish	175 228
JA	Japanese	852 087	HE	Hebrew	144 959
SV	Swedish	780 872	HI	Hindi	96 869
PT	Portuguese	758 227	EL	Greek	82 563
ZH	Chinese	663 485	TH	Thai	78 953

About 17M wikipages considered
(March '13)

WIKIPEDIA
The Free Encyclopedia

Each Wikipedia edition is treated as a complex network



Wikipedia PageRanking of World Universities WRWU

1st	University of Cambridge
2nd	University of Oxford
3rd	Harvard University
4th	Columbia University
5th	Princeton University
6th	MIT
7th	University of Chicago
8th	Stanford University
9th	Yale University
10th	University of California, Berkeley

90% overlap
between top 10s
WRWU and ARWU

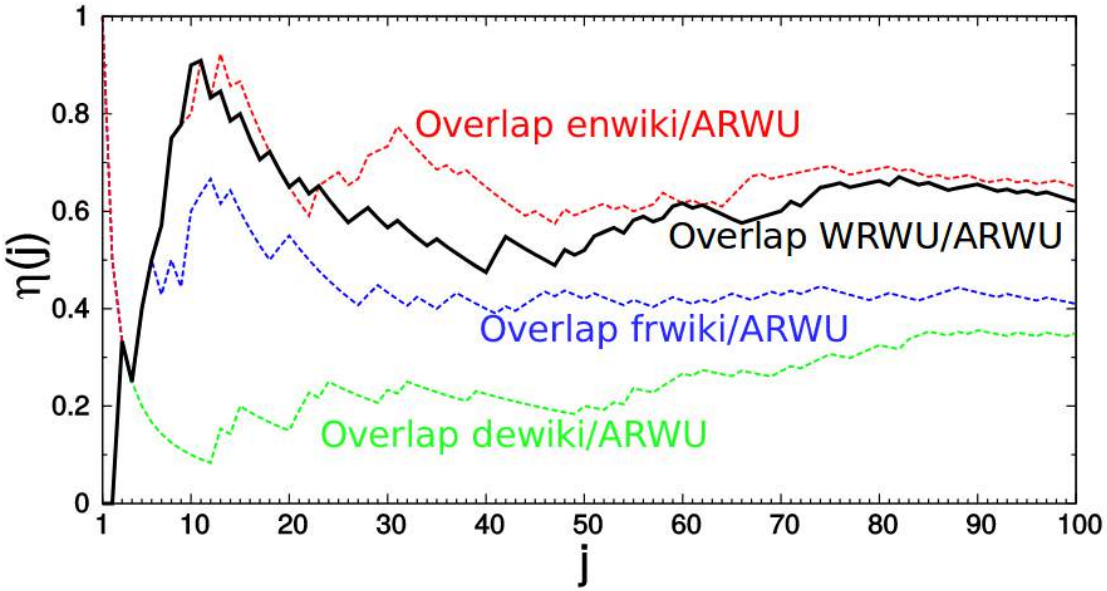
60% overlap
between top 100s
WRWU and ARWU

Academic Ranking of World Universities ARWU ("Shanghai ranking" 2013)

1st	Harvard University (-2)
2nd	Stanford University (-6)
3rd	University of California, Berkeley (-7)
4th	MIT (-2)
5th	University of Cambridge (+4)
6th	California Institute of Technology (-22)
7th	Princeton University (+2)
8th	Columbia University (+4)
9th	University of Chicago (+2)
10th	University of Oxford (+8)

Oxbridge at the top of WRWU
followed by **US major universities**

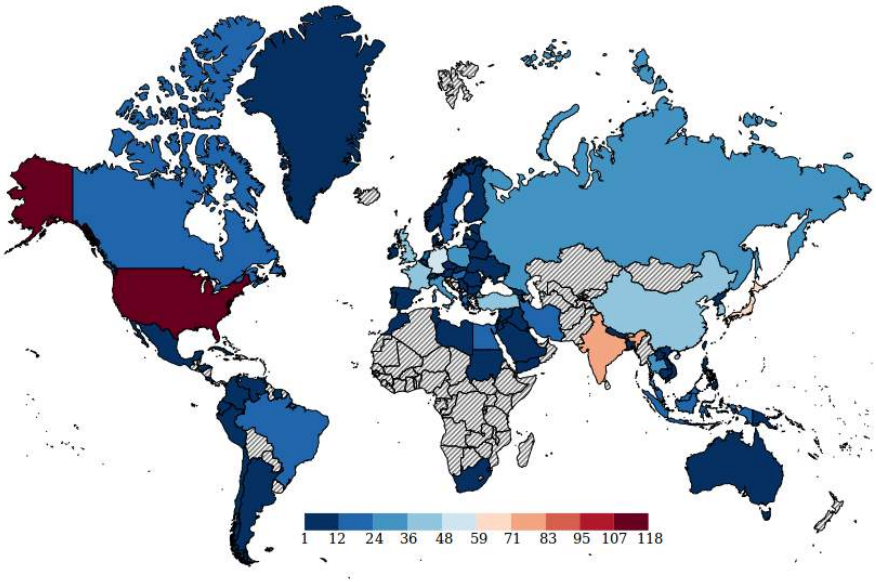
Marketwatch, December 10, '15



Overlap $\eta(j) = j_c/j$ is the ratio of common universities among the first j

Definitively, as ARWU, WRWU measures academic excellence, but not only ...

Geographical distribution of universities in WRWU



"Newcomers" in top 100

XIth century
25 11 University of Bologna

XIIIth century
97 13 University of Coimbra
69 13 University of Padua

XIVth century
33 14 Charles University in Prague
65 14 Jagiellonian University
51 14 Sapienza University of Rome
21 14 University of Vienna

XVth century
26 15 Leipzig University
59 15 University of Glasgow
92 15 University of St Andrews
64 15 University of Tübingen

XVIth century
90 16 Martin Luther University of Halle-Wittenberg
80 16 Trinity College, Dublin
75 16 University of Jena

XVIIth century
32 17 Lund University
93 17 University of Amsterdam
83 17 University of Tartu

XVIIIth century
38 18 École Polytechnique
56 18 Georgetown University
66 18 Saint Petersburg State University
22 18 University of Göttingen
99 18 University of Wrocław

XIXth century
11 19 Humboldt University of Berlin
98 19 Indiana University
76 19 Keio University
23 19 London School of Economics
61 19 Peking University
39 19 University of Bonn
95 19 University of Notre Dame
45 19 University of Virginia
86 19 University of Warsaw
71 19 Waseda University

XXth century
43 20 Al-Azhar University
67 20 Free University of Berlin
85 20 Institut Polytechnique des Sciences Avancées
96 20 Technical University of Berlin
91 20 Tsinghua University
100 20 University of Hamburg

These universities are important by their historical, social, or regional impact.

Wikipedia Ranking of World Universities using PageRank algorithm
WPRWU

Theta_PR = Theta PageRank score / Na = Number of appearances in the 24 Wikipedia editions / CC = country code / LC = language code / FC = Foundation century
Universities are ranked by Theta PageRank score (descending order), then by number of appearance in the 24 Wikipedia editions (descending order) and then by foundation century (ascending order)
[Download dataset]

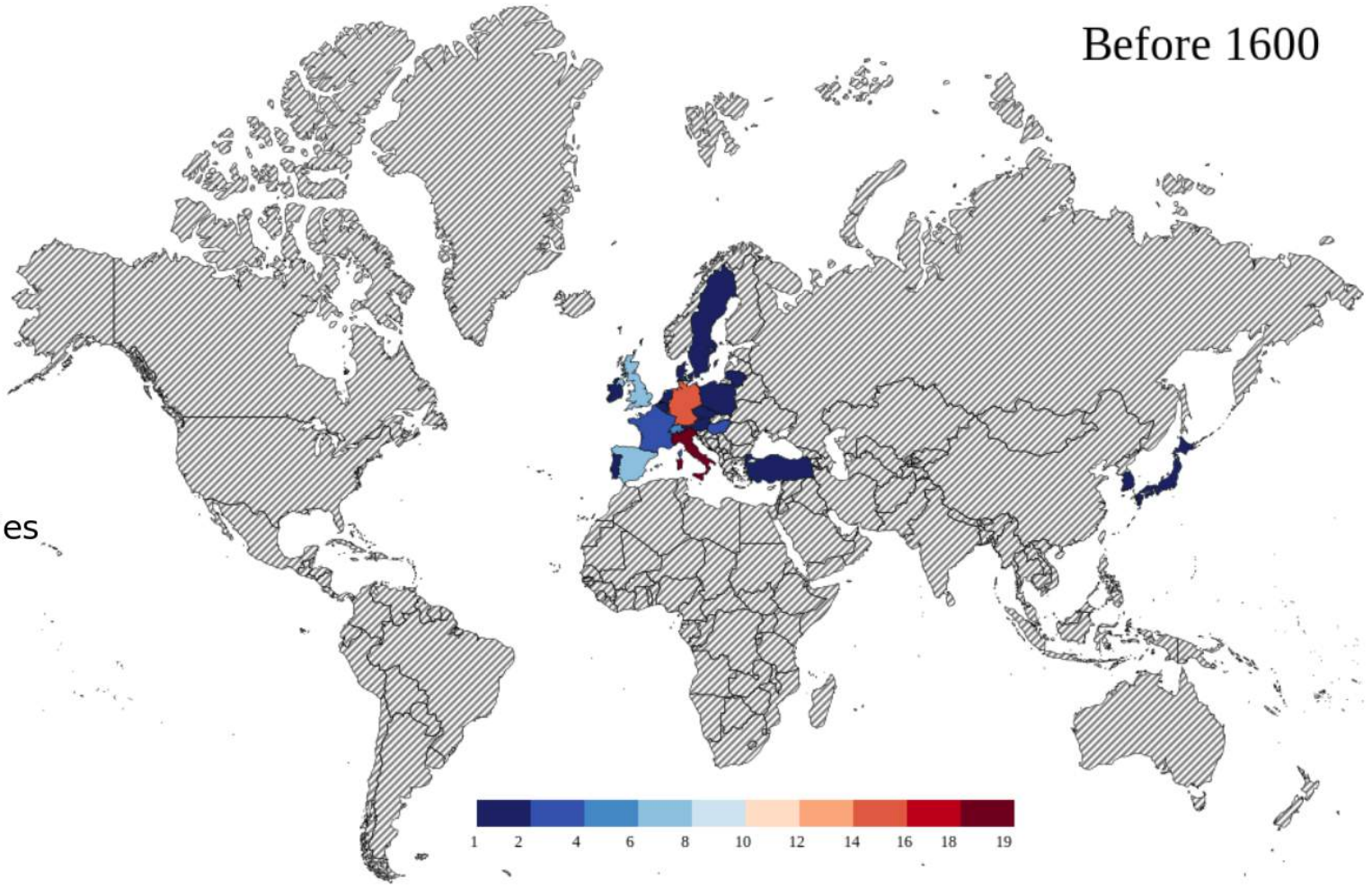
Universities in Top 5 were founded before XIXth century

Rank	Theta_PR	Na	University	CC	LC	FC
1	2272	24	University of Cambridge	UK	EN	13
2	2247	24	University of Oxford	UK	EN	11
3	2112	22	Harvard University	US	EN	17
4	2025	23	Columbia University	US	EN	18
5	1887	23	Princeton University	US	EN	18
6	1869	21	Massachusetts Institute of Technology	US	EN	19
7	1783	22	University of Chicago	US	EN	19
8	1765	21	Stanford University	US	EN	19
9	1716	20	Yale University	US	EN	18
10	1557	19	University of California, Berkeley	US	EN	19
11	1531	21	Humboldt University of Berlin	DE	DE	19
12	1481	22	Cornell University	US	EN	19
13	1351	20	University of Pennsylvania	US	EN	18
14	1285	20	University of London*	UK	EN	19
15	1224	19	Uppsala University	SE	SV	15
16	1209	20	University of Edinburgh	UK	EN	16
17	1195	20	Heidelberg University	DE	DE	14
18	1193	18	University of California, Los Angeles	US	EN	19
19	1171	20	New York University	US	EN	19
20	1131	18	University of Michigan	US	EN	19
21	1119	19	Johns Hopkins University	US	EN	19
22	1113	19	University of Vienna	AT	DE	14
23	1099	18	University of Göttingen	DE	DE	18
24	1030	16	London School of Economics	UK	EN	19
25	990	19	Moscow State University	RU	RU	18
26	974	19	University of Bologna	IT	IT	11
27	948	18	Leipzig University	DE	DE	15
28	928	15	California Institute of Technology	US	EN	19
29	911	18	Ludwig Maximilian University of Munich	DE	DE	15
30	764	15	University of Southern California	US	EN	19
31	752	17	University of Tokyo	JP	JA	19
32	743	15	Leiden University	NL	NL	16
33	707	11	Lund University	SE	SV	17
34	680	13	Charles University in Prague	CZ	WR	14
35	668	12	University College London	UK	EN	19
36	577	11	University of Copenhagen	DK	DA	15
37	576	11	École Normale Supérieure	FR	FR	18
38	570	14	University of Manchester	UK	EN	19
39	556	13	École Polytechnique	FR	FR	18
40	538	14	University of Bonn	DE	DE	19
41	523	11	University of Texas at Austin	US	EN	19
42	519	15	Duke University	US	EN	19
43	507	15	Carnegie Mellon University	US	EN	19
44	505	9	Al-Azhar University	EG	AR	20
45	490	10	University of Helsinki	FI	WR	17
46	487	15	University of Virginia	US	EN	19
47	483	12	Hebrew University of Jerusalem	IL	HE	20
48	470	12	University of Toronto	CA	EN	19
49	460	9	King's College London	UK	EN	16
50	450	9	Imperial College London	UK	EN	20
51	447	11	University of Illinois at Urbana–Champaign	US	EN	19
52	429	10	Sapienza University of Rome	IT	IT	14
53	428	8	ETH Zurich	CH	DE	19
54	426	12	University of Zurich	CH	DE	16
55	389	12	University of Washington	US	EN	19

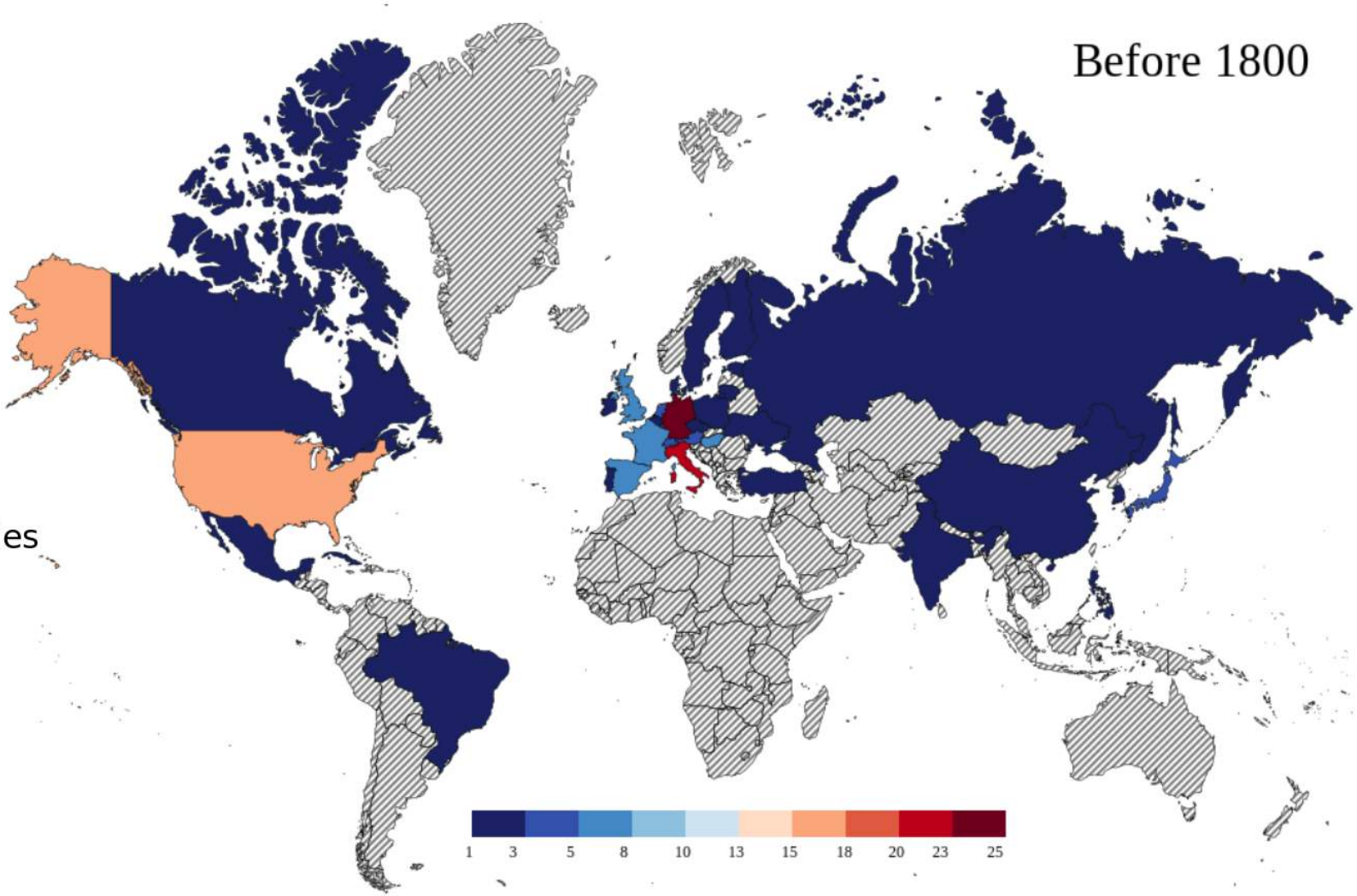
Universities in Top 43 were founded before XXth century

Quite rigid club of first universities "not willing" to accept new members

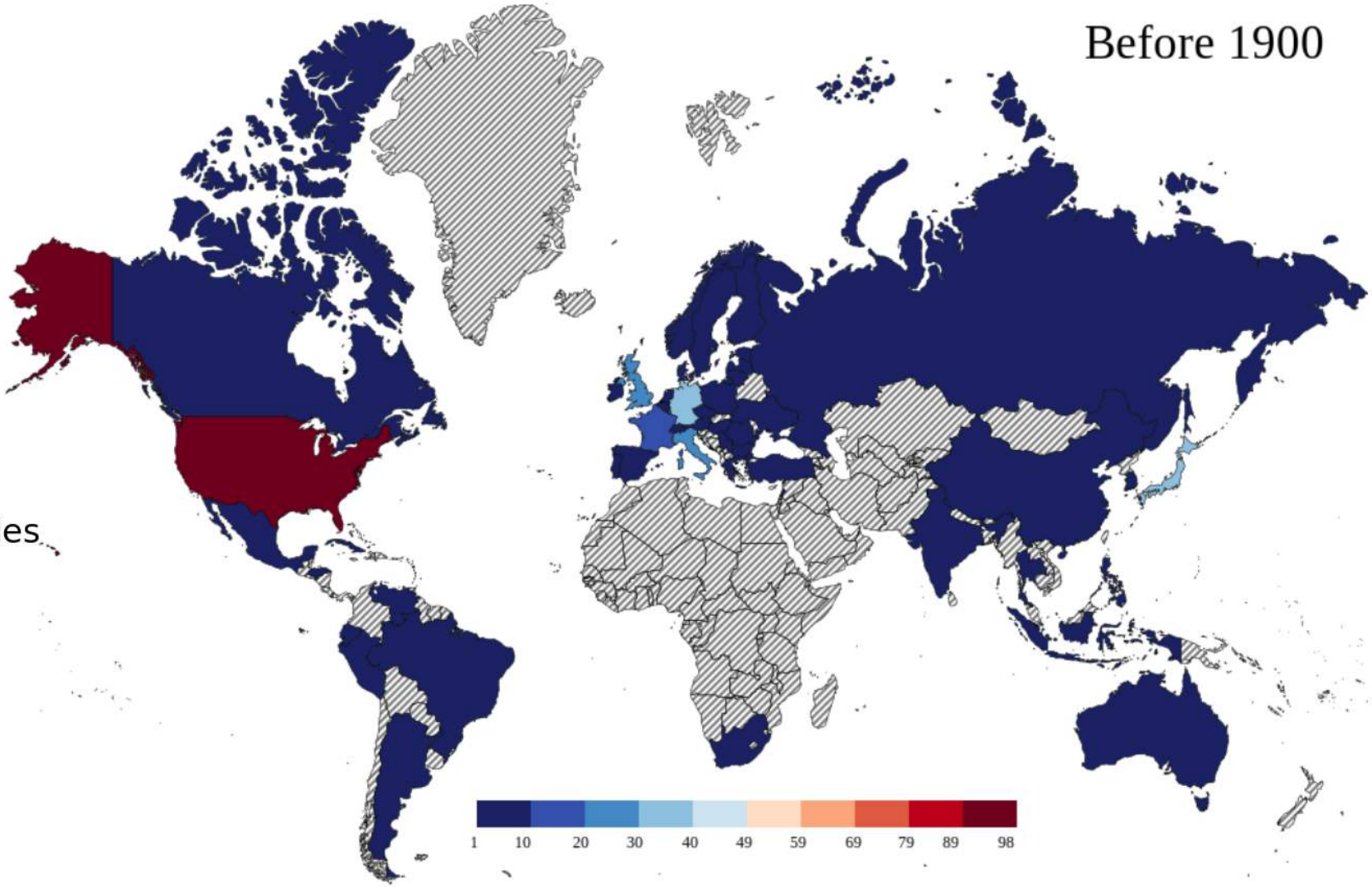
Dominance of European universities



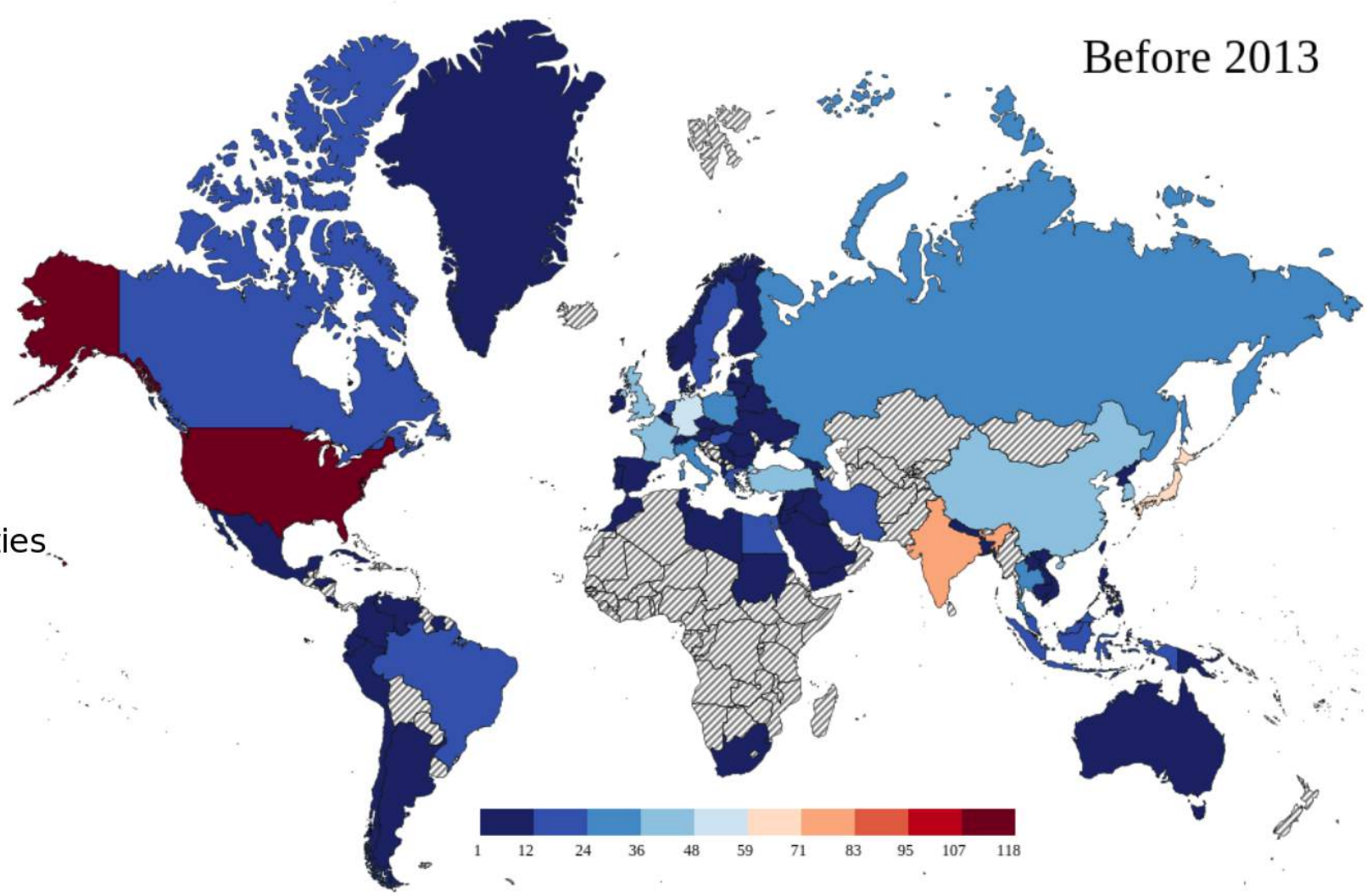
Emergence of US universities



Dominance of US universities

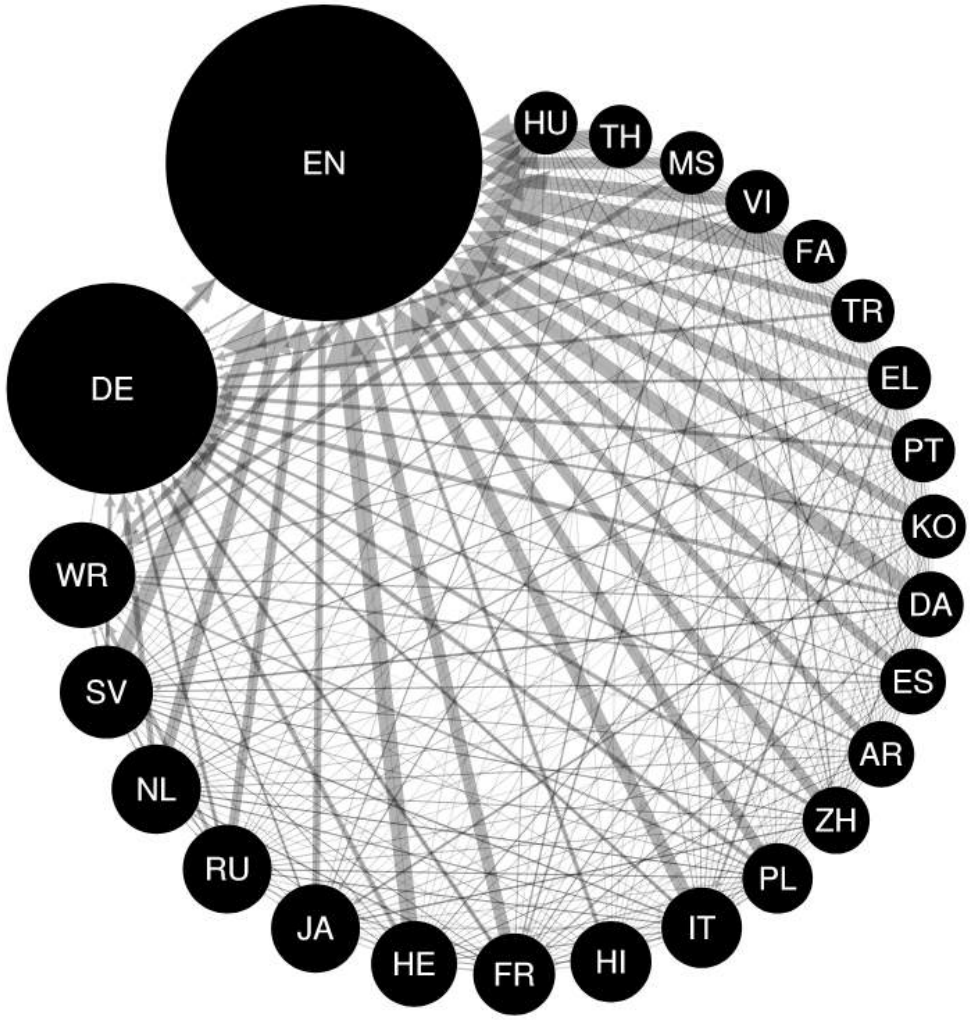


Emergence of Asian universities



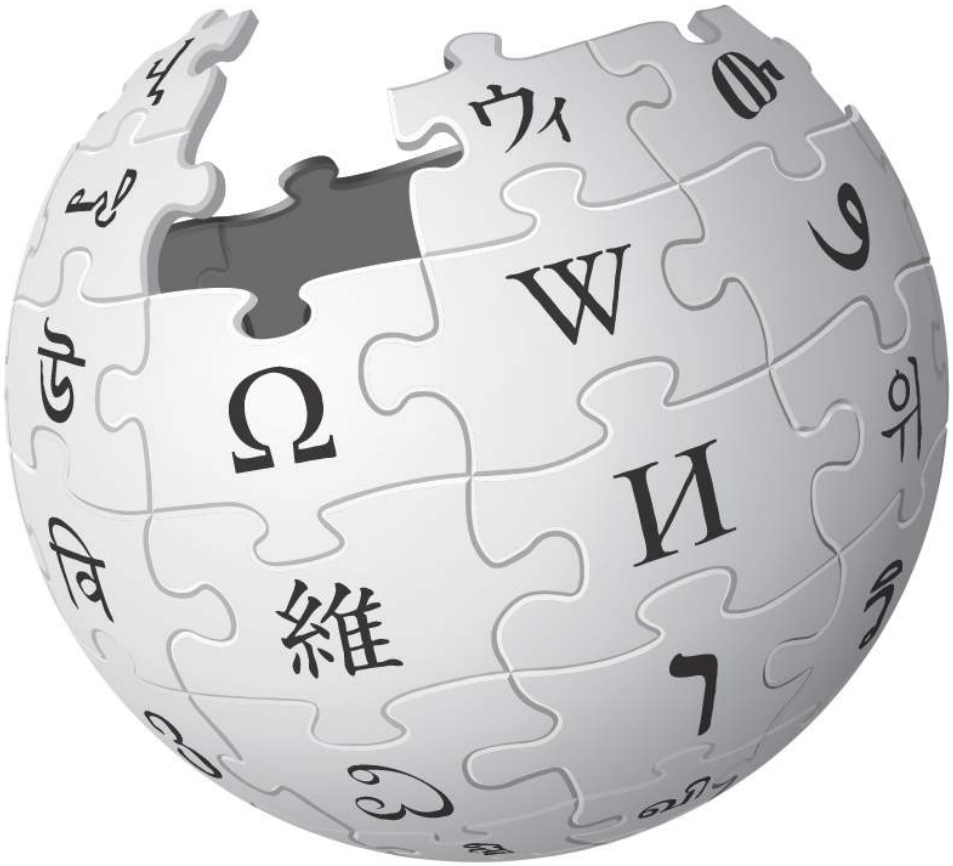
Network of cultures

An arrow points from a culture A to a culture B,
the width of the arrow is proportional to the number of university of culture B
appearing in the ranking of culture A



Wikipedia Ranking of World Universities

Conclusion



WRWU is **free from any cultural preferences since** :

- it takes into account many cultural points of view as we use all human knowledge contained in 24 Wikipedia language editions (17 millions Wiki articles)
- these cultural points of view are treated on equal footing with the same statistical analysis (PageRank, CheiRank, 2DRank)

WRWU measures **academic excellence** (top 10 and top 100 are similar to ARWU) but also **historic, social, or regional importance** of universities.

WRWU can be considered as **complementary** to already existing rankings such as ARWU, but **in fact it encodes already all existing rankings** since Wikipedia contains information on it.

Universal ranking ?

J. Lages, A. Patt, D. L. Shepelyansky, *Wikipedia Ranking of World Universities,* The European Physical Journal B (2016) 89: 69

Is your alma mater well ranked ?
Complete ranking at:

http://perso.utinam.cnrs.fr/~lages/datasets/WRWU/

Reduced Google matrix

Consider a network with $N \gg 1$ nodes.
Consider a sub-network (a community) of $N_r \ll N$ nodes. Google matrix of the N size network and the associated PageRank vector can be written

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_{rr} & \mathbf{G}_{rs} \\ \mathbf{G}_{sr} & \mathbf{G}_{ss} \end{pmatrix}, \qquad \mathbf{P} = \begin{pmatrix} \mathbf{P}_r \\ \mathbf{P}_s \end{pmatrix}$$

$$\mathbf{GP} = \mathbf{P}$$

We define the reduced Google matrix $\mathbf{G}_R$ associated to the size N_r community such as

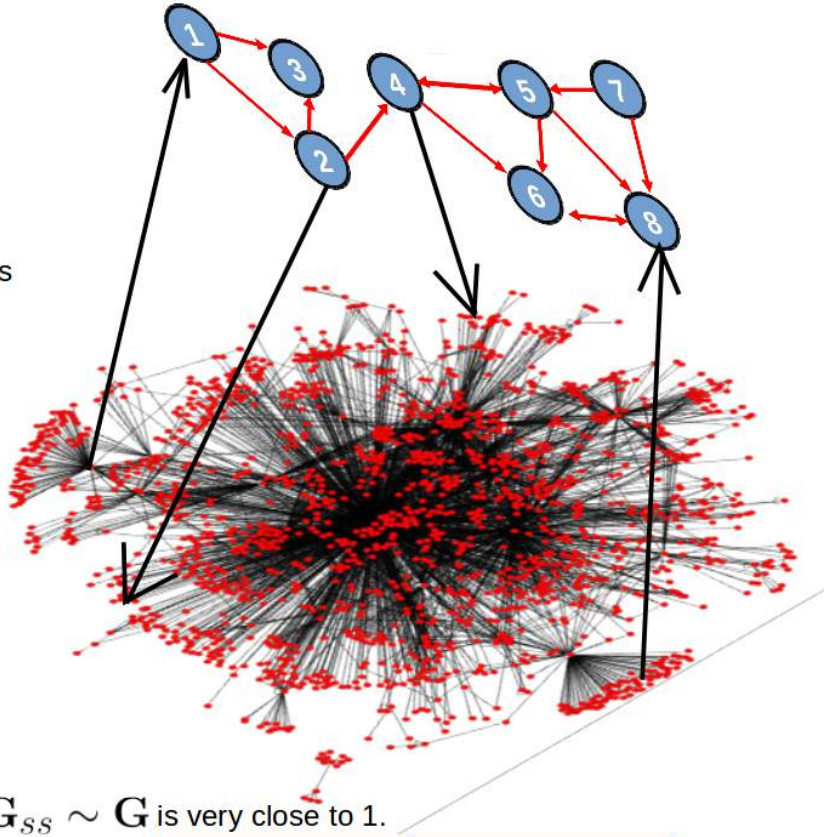
$$\mathbf{G}_R \mathbf{P}_r = \mathbf{P}_r$$

The reduced Google matrix can be written

$$\mathbf{G}_R = \mathbf{G}_{rr} + \mathbf{G}_{rs}\left(\mathbf{1} - \mathbf{G}_{ss}\right)^{-1}\mathbf{G}_{sr}$$

Contribution from direct links

Contribution from indirect links (scattering term)



$$\left(\mathbf{1} - \mathbf{G}_{ss}\right)^{-1} = \sum_{l=0}^{\infty} \mathbf{G}_{ss}^{l}$$

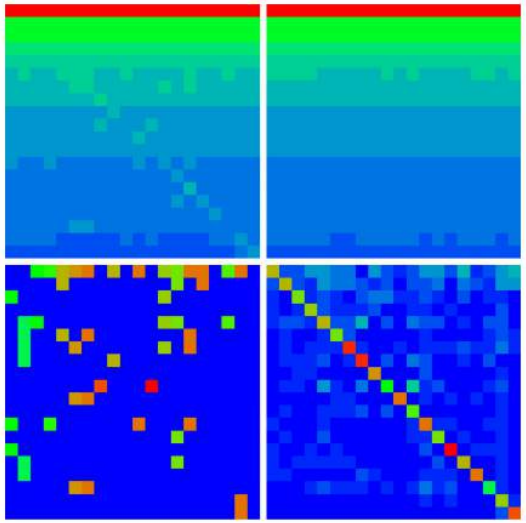
Very slow convergence since the eigenvalue λ_c of $\mathbf{G}_{ss} \sim \mathbf{G}$ is very close to 1.

$$\mathbf{G}_R = \mathbf{G}_{rr} + \mathbf{G}_{pr} + \mathbf{G}_{qr}$$

Contribution from direct links

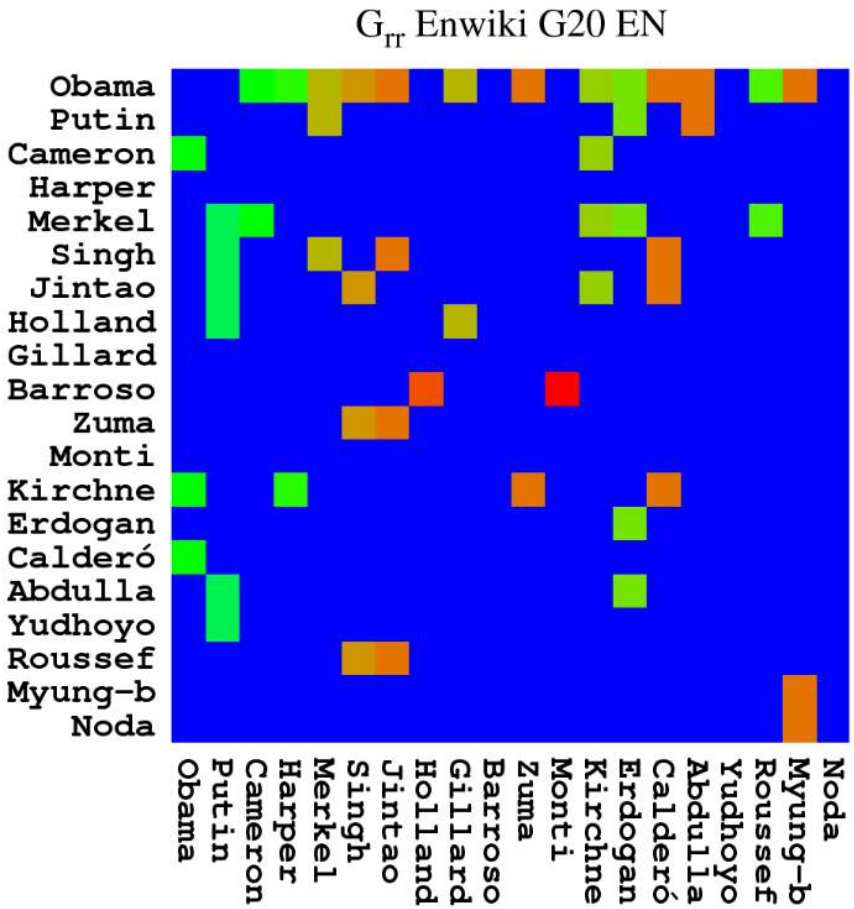
Contribution from hidden links

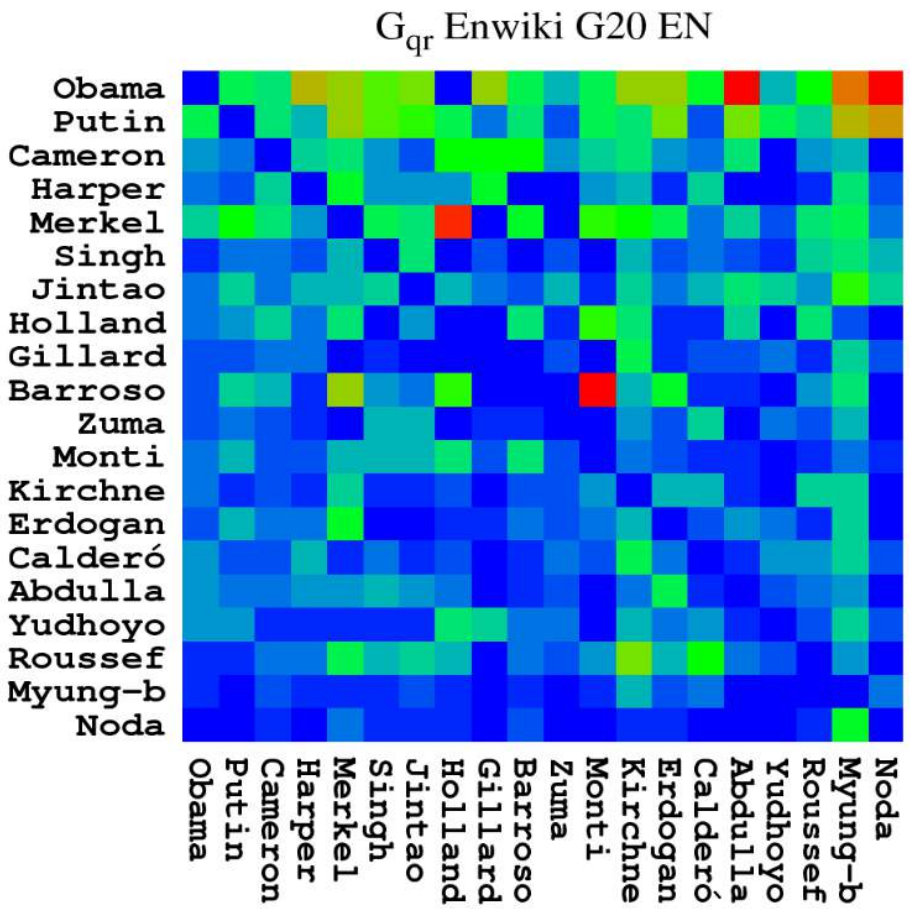
Contribution from « PageRank »



Wikipedia mining of hidden links between political leaders

2013 Wikipedia edition





$$\mathbf{G}_R = \mathbf{G}_{rr} + \mathbf{G}_{pr} + \mathbf{G}_{qr}$$

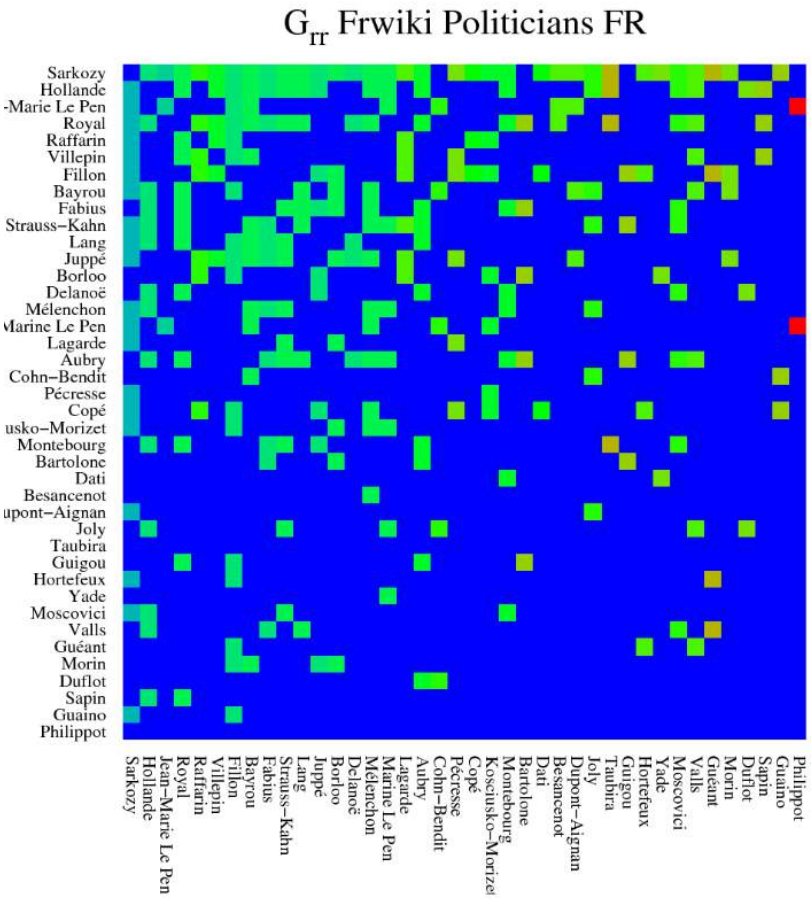
Contribution from direct links

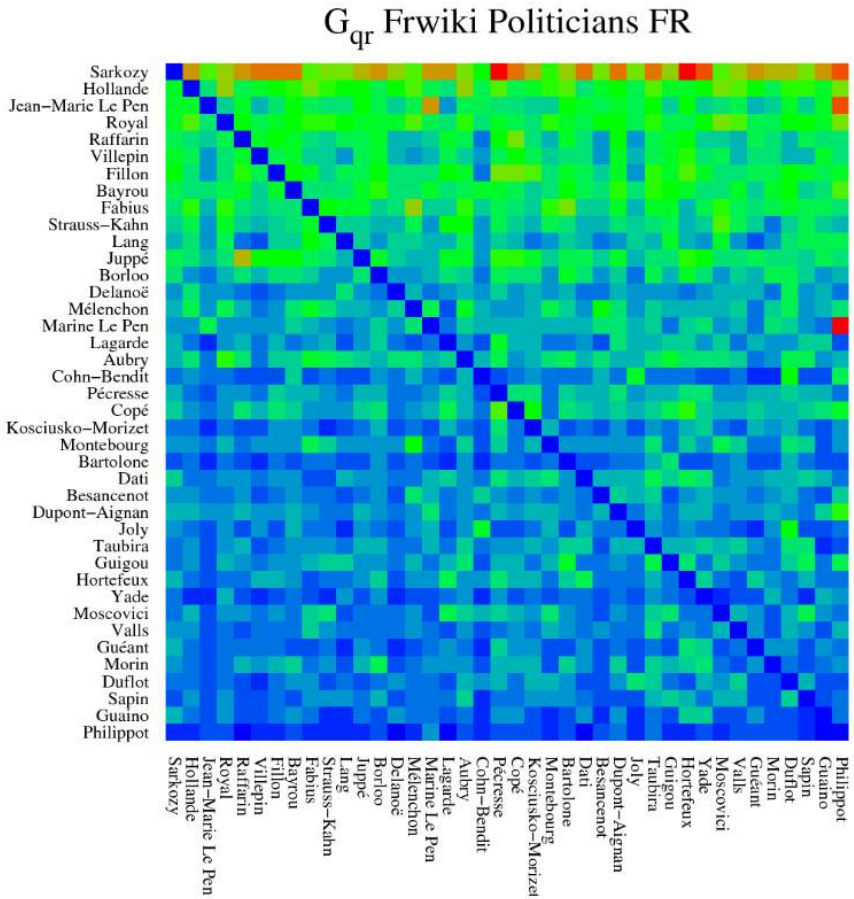
Contribution from hidden links

Frahm, Jaffrès-Runser, Shepelyansky, Eur. Phys. J. B (2016) 89: 269

Wikipedia mining of hidden links between political leaders

2013 Wikipedia edition





$$\mathbf{G}_R = \mathbf{G}_{rr} + \mathbf{G}_{pr} + \mathbf{G}_{qr}$$

Contribution from direct links

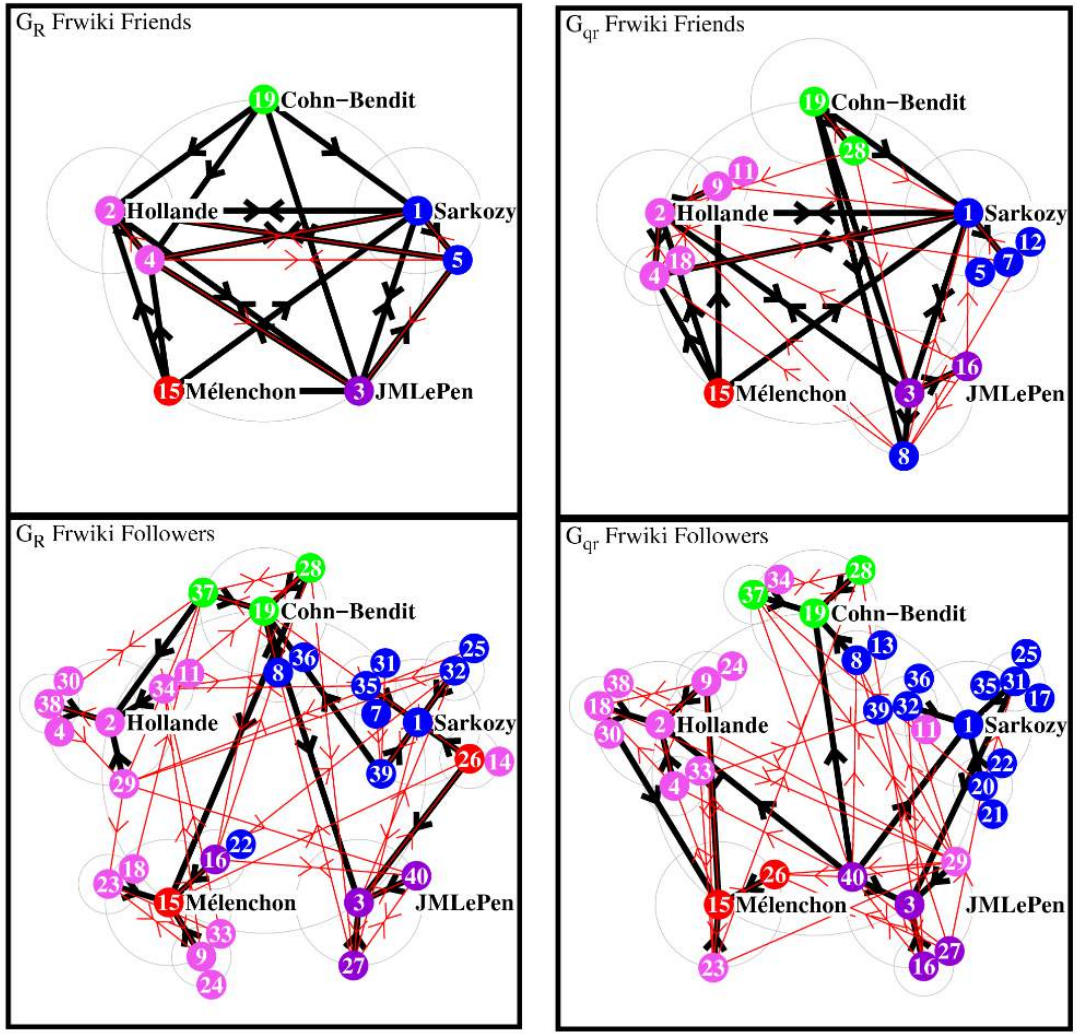
Contribution from hidden links

Frahm, Jaffrès-Runser, Shepelyansky, Eur. Phys. J. B (2016) 89: 269

Wikipedia mining of hidden links between political leaders

Circles of influence

2013 Wikipedia edition



Names (FR)	K
Nicolas Sarkozy CB	1
François Hollande CM	2
Jean-Marie Le Pen CV	3
Ségolène Royal CM	4
Jean-Pierre Raffarin CB	5
Dominique de Villepin CB	6
François Fillon CB	7
François Bayrou CB	8
Laurent Fabius CM	9
Dominique Strauss-Kahn CM	10
Jack Lang CM	11
Alain Juppé CB	12
Jean-Louis Borloo CB	13
Bertrand Delanoë CM	14
Jean-Luc Mélenchon CR	15
Marine Le Pen CV	16
Christine Lagarde CB	17
Martine Aubry CM	18
Daniel Cohn-Bendit CG	19
Valérie Pécresse CB	20
Jean-François Copé CB	21
Nathalie Kosciusko-Morizet CB	22
Arnaud Montebourg CM	23
Claude Bartolone CM	24
Rachida Dati CB	25
Olivier Besancenot CR	26
Nicolas Dupont-Aignan CV	27
Eva Joly CG	28
Christiane Taubira CM	29
Élisabeth Guigou CM	30
Brice Hortefeux CB	31
Rama Yade CB	32
Pierre Moscovici CM	33
Manuel Valls CM	34
Claude Guéant CB	35
Hervé Morin CB	36
Cécile Duflot CG	37
Michel Sapin CM	38
Henri Guaino CB	39
Florian Philippot CV	40

$$\mathbf{G}_R = \mathbf{G}_{rr} + \mathbf{G}_{pr} + \mathbf{G}_{qr}$$

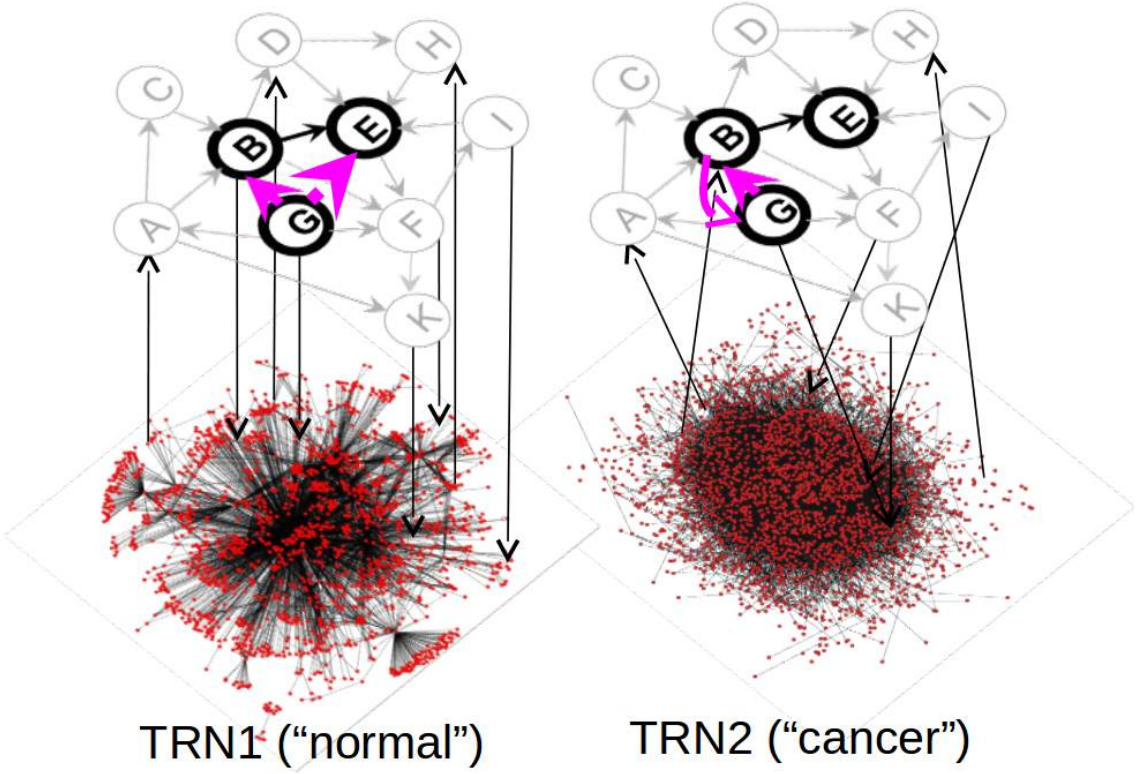
Contribution from direct links

Contribution from hidden links

Frahm, Jaffrès-Runser, Shepelyansky, Eur. Phys. J. B (2016) 89: 269

Comparing two TRN networks: e.g., "normal" vs "cancer"

(results presented at the 13th [BC]2 Basel Computational Biology Conference, 2017)



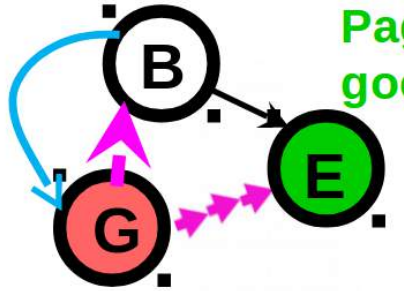
TRN1 (“normal”)

Normal B-Lymphocytes

TRN2 (“cancer”)

Leukemia cell line

indirect signaling rewiring



PageRank goes down

PageRank improves

hidden emergent oncogenic

hidden dissapearing in oncogenic network

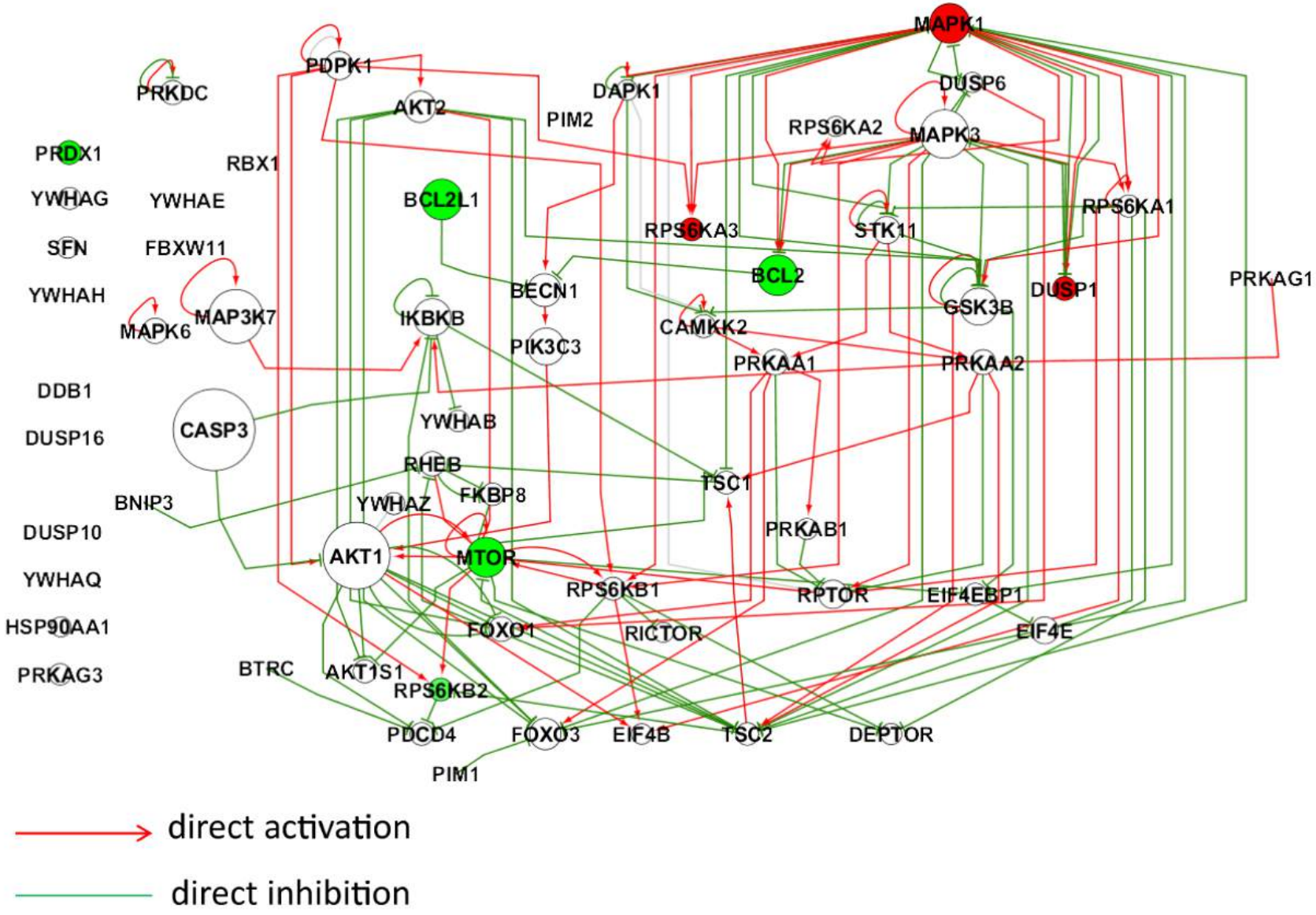
direct

indirect “hidden”

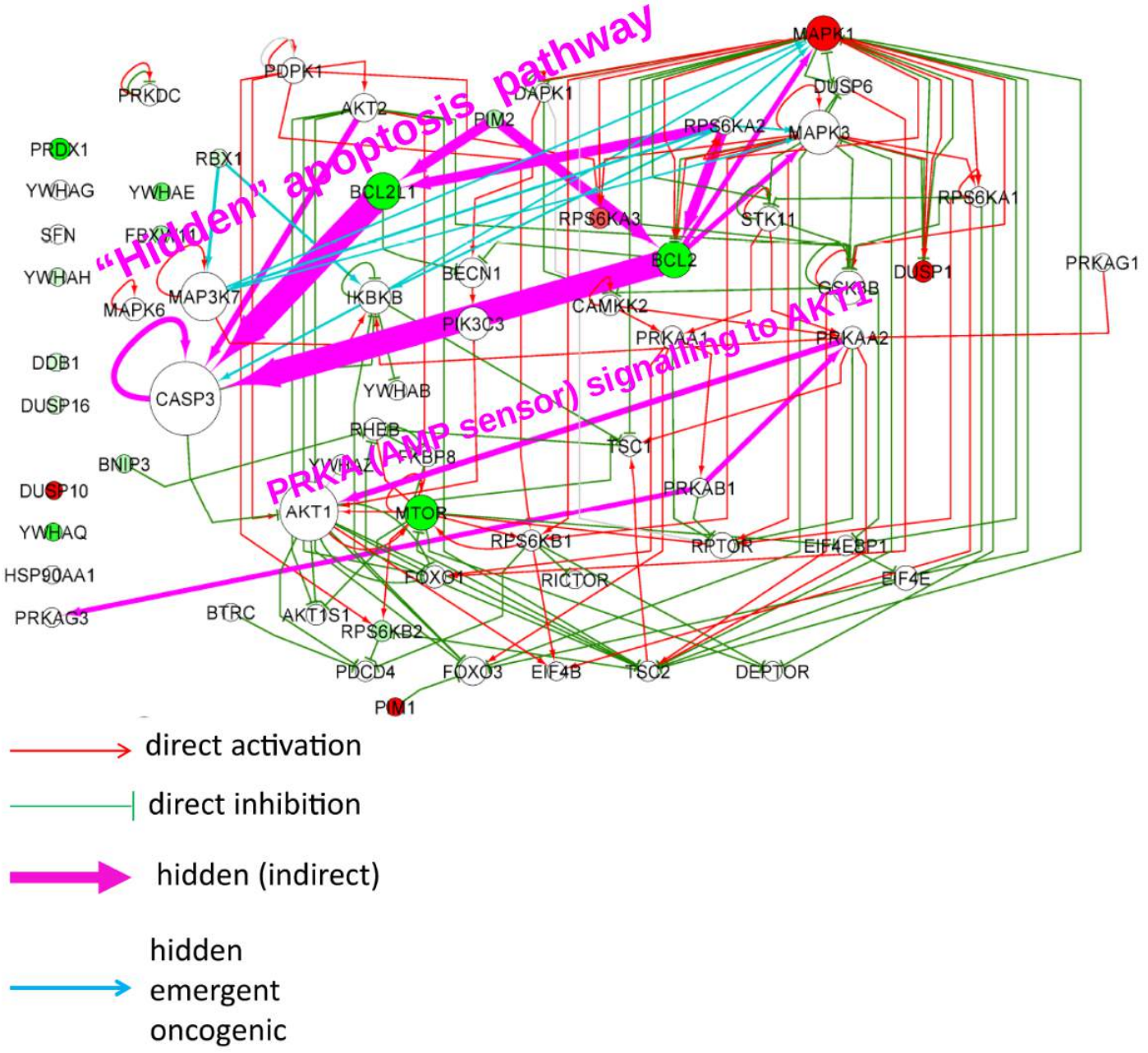
Reference:

J.L., D.S., A.Z., bioRxiv, https://doi.org/10.1101/096362 , soumis à PLOS ONE

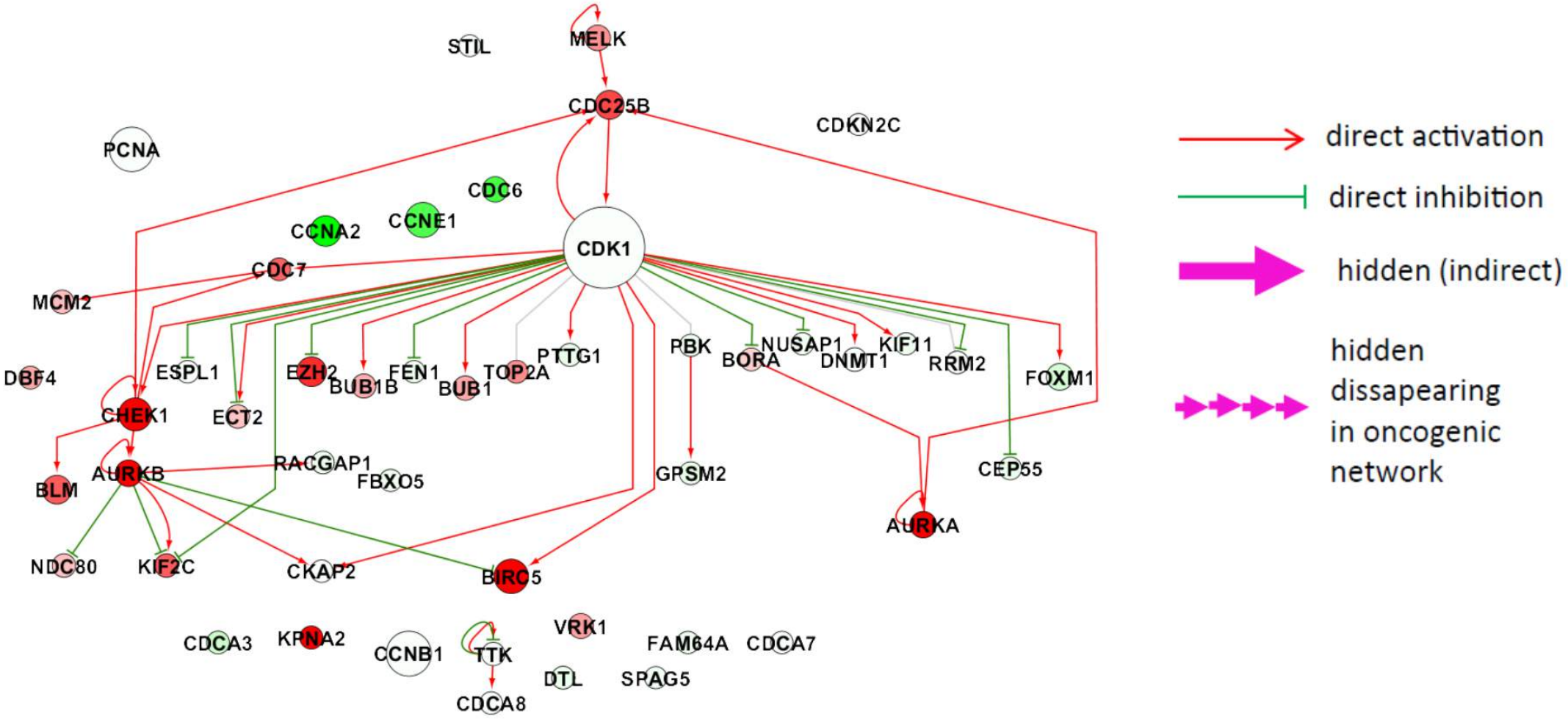
Inferring indirect (hidden) causal connections between AKT-mTOR pathway members



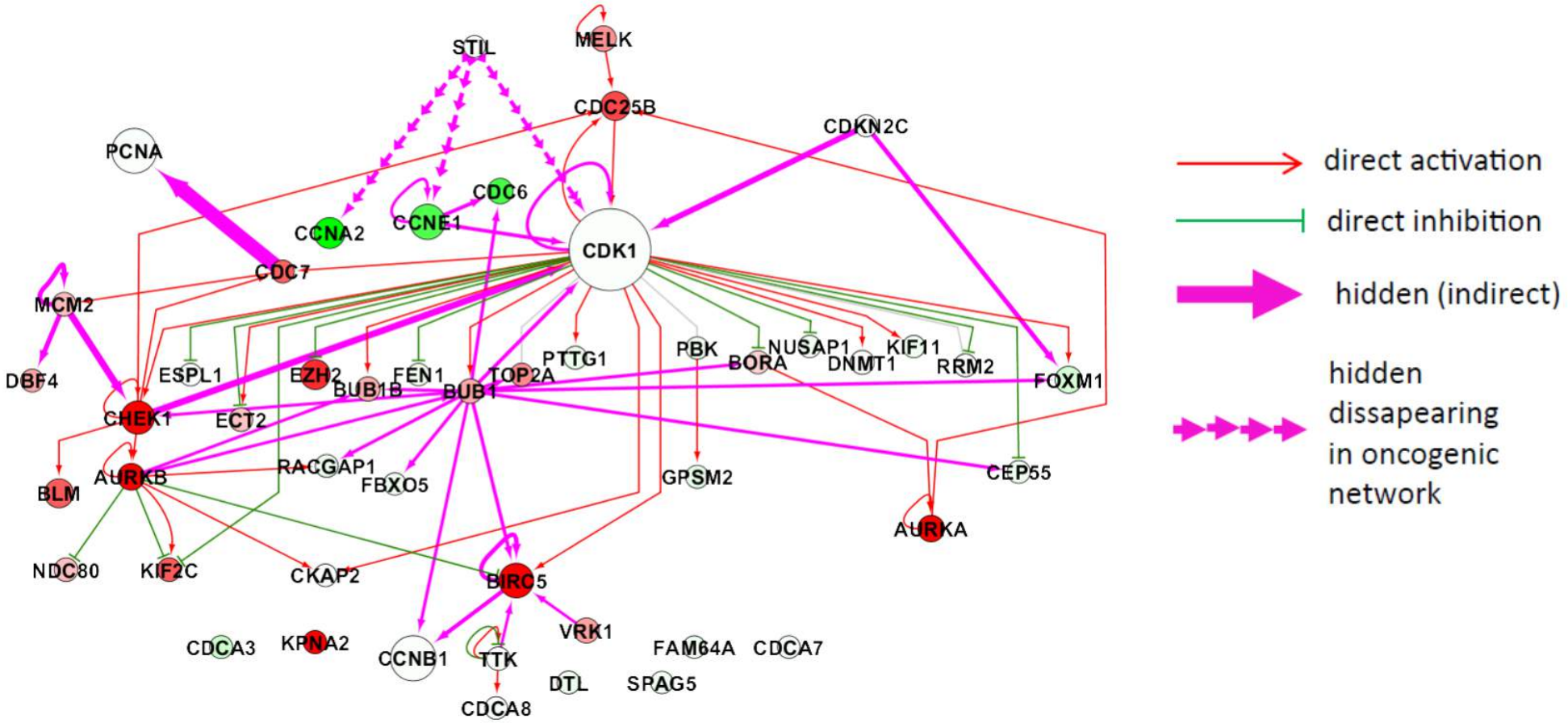
Inferring indirect (hidden) causal connections between pathway members



Genes of a proliferative signature resulted from pancancer transcriptomic analysis



Genes of a proliferative signature resulted from pancancer transcriptomic analysis



More genes are connected into the network
Emergence of a new “hidden” hub BUB1
Connection to PCNA (DNA replication and DNA repair)
Many cell cycle proteins improves in PageRank (AURK)
Connection between STIL (mitotic spindle checkpoint regulator) and CCNA2, CCNE1

Merci pour votre attention