SUNBELT 2019
French Connection

Google matrix analysis of Wikipedia networks

22th of June 2019, University of Québec Montréal
, CA-QCT

***Coquidé Célestin*[1], *Rollin Guillaume*[1],**
***Lages José*[1], *Shepelyansky Dima L.*[2]**

[1]*Institut UTINAM, university of Bourgogne Franche-Comté, Besançon , France*
[2]*Laboratoire de physique théorique, University of Paul Sabatier, Toulouse, France.*

http://perso.utinam.cnrs.fr/~lages/apex/

Institut UTINAM
cnrs
UBFC
UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
APEX
Analyse Physique des résEaux compleXes
RÉGION BOURGOGNE FRANCHE COMTÉ
UFC
UNIVERSITÉ DE FRANCHE-COMTÉ

Outline

1) Introduction
2) Wikipedia, a directed network
3) Google matrix and PageRank
4) Wikipedia ranking of world universities 2017 (WRWU17)
5) Reduced google matrix (REGOMAX)
6) World influence and Interactions of Universities
7) Conclusion

1) Introduction

- **Wikipedia** is a **complete encyclopedia** **47,803,467 articles**
- Worldwide improved with **294** different language **editions**
- Very heterogeneous **knowledge** base

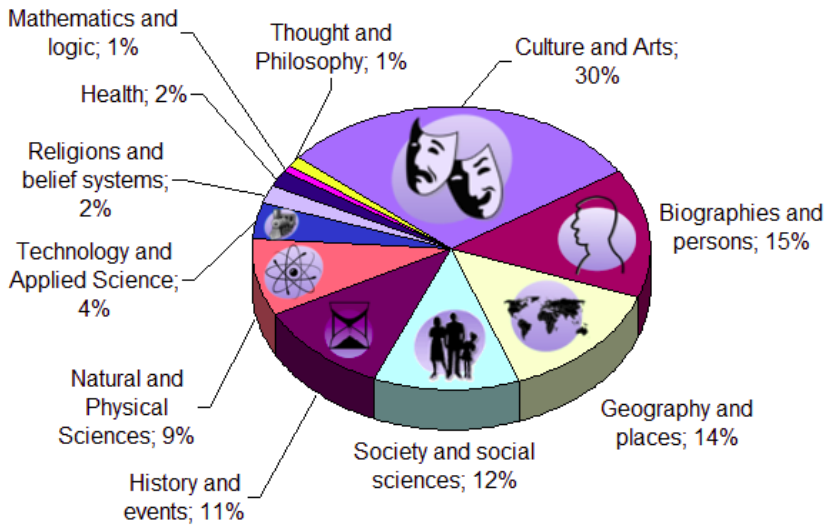


Fig. 1 Distribution of wikipedia content by subject in 2008 (Kittur et al. 2009)

2) Wikipedia, a directed network

Nodes

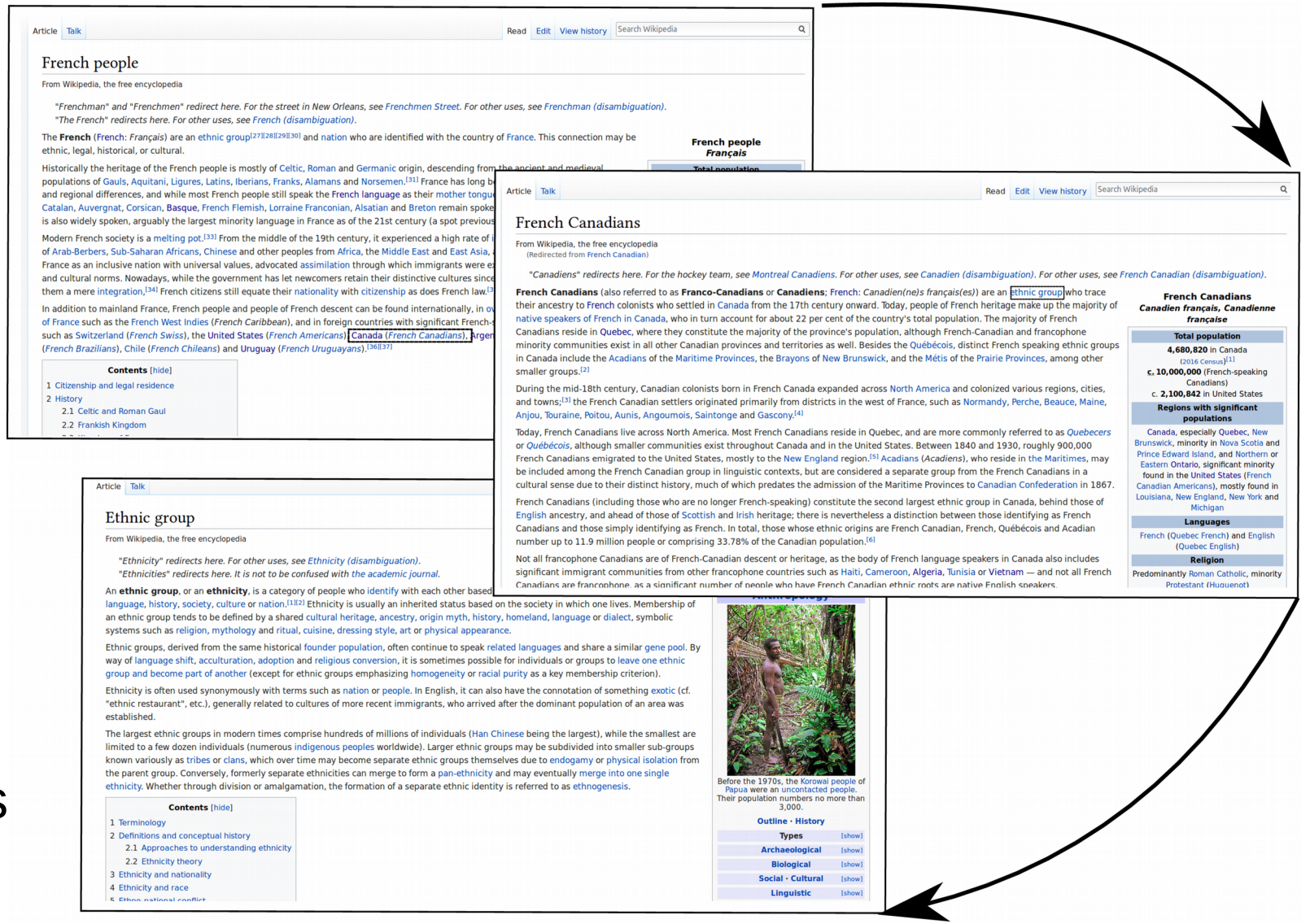
=

articles

Links

=

hyperlinks



2) Wikipedia, a directed network

- Different **cultural sights**
- **Various research topics** can be studied using **WIKINET** (geopolitic, geography, history, economy …)
- **Data updated regularly** and **easily available**.

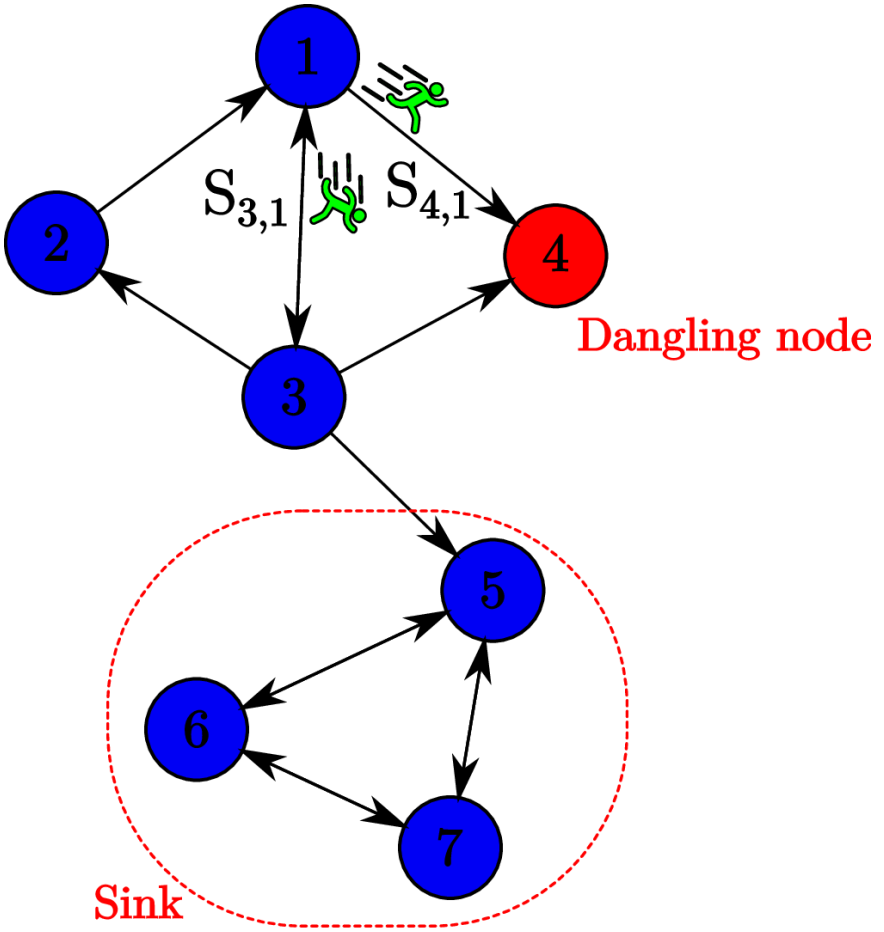
How can we get information from such a network?

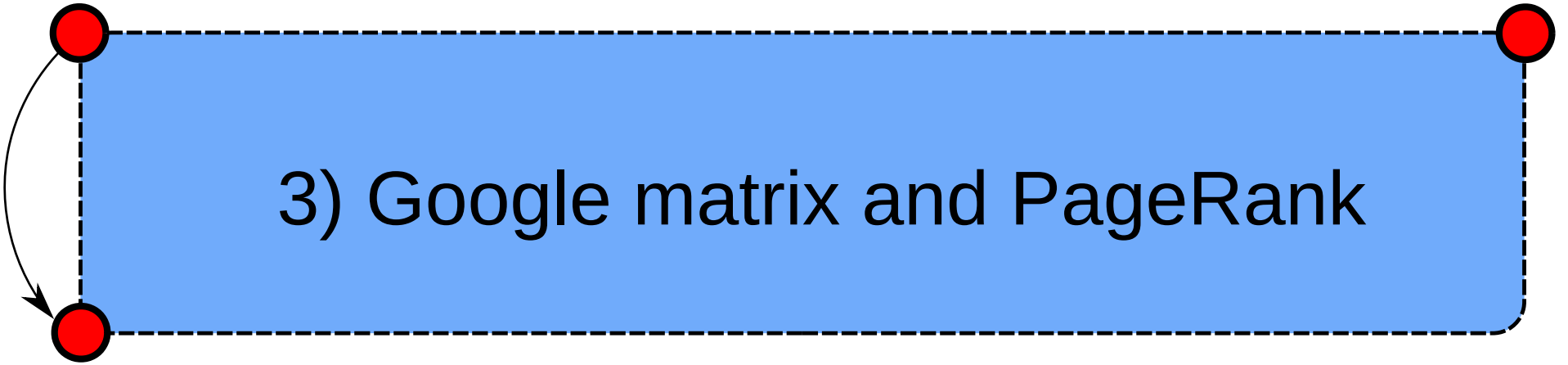
Edition	Language	*N*	Edition	Language	*N*
EN	English	5416537	ZH	Chinese	939625
SV	Swedish	3786455	FA	Persian	539926
DE	German	2057898	AR	Arabic	519714
NL	Dutch	1900222	HU	Hungarian	409297
FR	French	1866546	KO	Korean	380086
RU	Russian	1391225	TR	Turkish	291873
IT	Italian	1353276	MS	Malaysian	289234
ES	Spanish	1287834	DA	Danish	225523
PL	Polish	1219733	HE	Hebrew	205411
VI	Vietnamese	1155932	EL	Greek	130429
JP	Japanese	1058950	HI	Hindi	121503
PT	Portuguese	967162	TH	Thai	116495

Tab. 1 Set of the 24 most complete language editions of WIKI2017. Only 71818 articles for the ‘Encyclopédie’.

3) Google matrix and PageRank

- **Modeling** a **random walk** through a **directed network**
- **Stochastic** matrix (**S**)
  - Transition **probabilities**
  - **Dangling node** fixed
- **Google** matrix (**G**) (Brin and Page 1998)
  - **Sink** effect **avoiding**
  - **Leading eigenvalue** degeneracy = 1





3) Google matrix and PageRank

- **PageRank** (P) = **leading eigenvector** such that (***GP=P***) is a

  => **Steady-state**

  => The **highest component** for the **most reachable node**

  => **Efficiency** of **ingoing** link = **Importance**

- **Inverting** links **direction** => *G** => **CheiRank**

  => **Efficiency** of **outgoing** link = **communicability**

- We can **rank** nodes by **PageRank/CheiRank** order

Is the ranking of wikipedia articles efficient?

4) Wikipedia ranking of world universities 2017 (WRWU17)

- Shanghai ranking method = ad hoc tool with human rules

  => economical bias, number of articles in nature, nobel prize

- We use **24 different Language editions** of **wikipedia 2017** to get a multicultural based ranking tool.

  => 24 top 100 of articles belonging to universities

- From 24 edition (*E)*, we compute theta score

$$\theta_U = \sum_{E=1}^{24} (101 - R_{U,E})$$

Rank	Θ_{PR}	N_a	University	CC	LC	FC
1st	2281	24	University of Oxford	UK	EN	11
2nd	2278	24	University of Cambridge	UK	EN	13
3rd	2277	24	Harvard University	US	EN	17
4th	2099	24	Columbia University	US	EN	18
5th	1959	23	Yale University	US	EN	18
6th	1917	24	University of Chicago	US	EN	19
7th	1858	23	Princeton University	US	EN	18
8th	1825	21	Stanford University	US	EN	19
9th	1804	21	Massachusetts Institute of Technology	US	EN	19
10th	1693	20	University of California, Berkeley	US	EN	19

Tab. 2: The top 10 universities from **WRWU17** *completer ranking (1011 Univ) at:* *http://perso.utinam.cnrs.fr/~lages/datasets/WRWU17/* **Check the position of your Alma Mater !!**

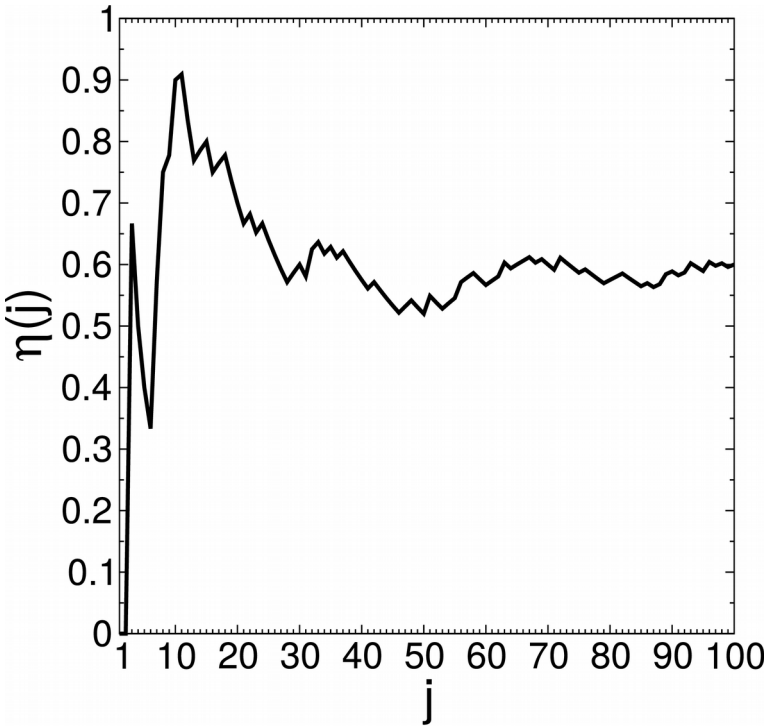


Fig. 2: Overlap between WRWU17 and Shanghai Ranking.

5) Reduced google matrix (REGOMAX)

- **REGOMAX** allows us to **point out hidden links** between nodes of interest (Frahm and Shepelyansky 2016, Frahm et al. 2016)

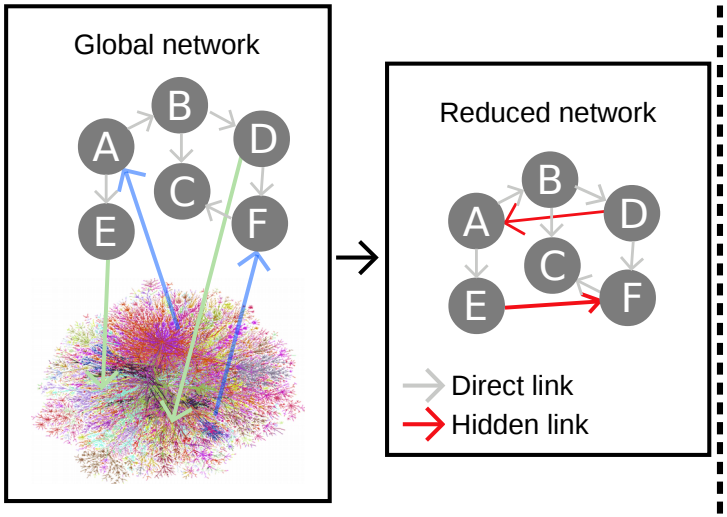
Transition between the N_r nodes

Transitions from the other N-N_r nodes toward the N_r nodes

$$G = \begin{pmatrix} G_{rr} & G_{rs} \\ G_{sr} & G_{ss} \end{pmatrix}$$

Transitions from the N_r nodes toward the other N-N_r nodes

Transitions between the other N-N_r nodes



$$G_R = G_{rr} + G_{pr} + G_{qr}$$

Direct links Contribution from **Hidden** links **«PageRank »**

6) World influence and Interactions of Universities

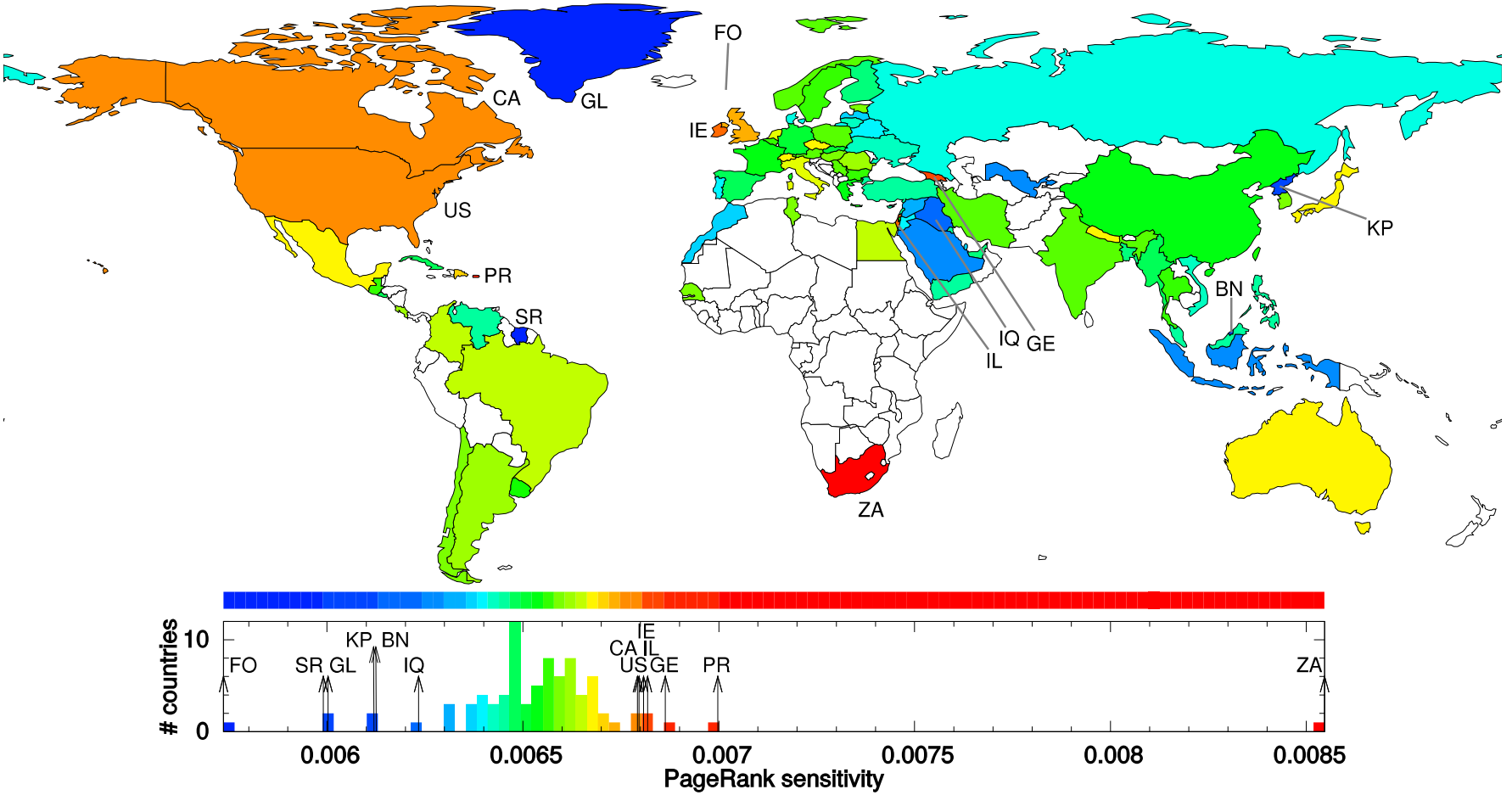


Fig. 3: Worldwide influence of Harvard. Color code represent the logarithm derivative of countries PageRank value when link coming from Harvard has been boosted.

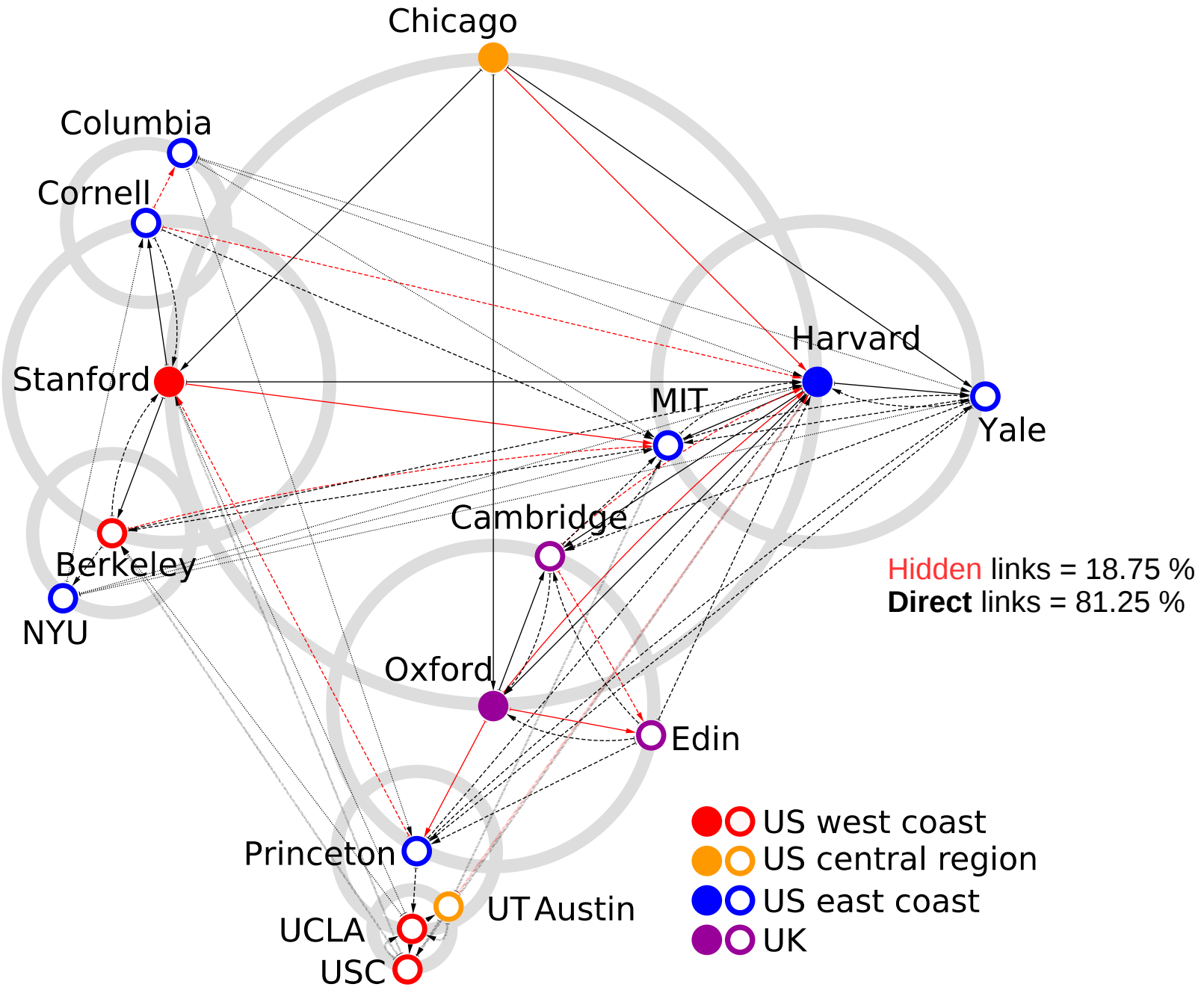


Fig. 4: ENWIKI17 top 20 univ. reduced network: **Direct** and hidden links, solid lines (1st level), dashed lines (2nd), doted lines (3rd) and "\" symbol lines (4+)

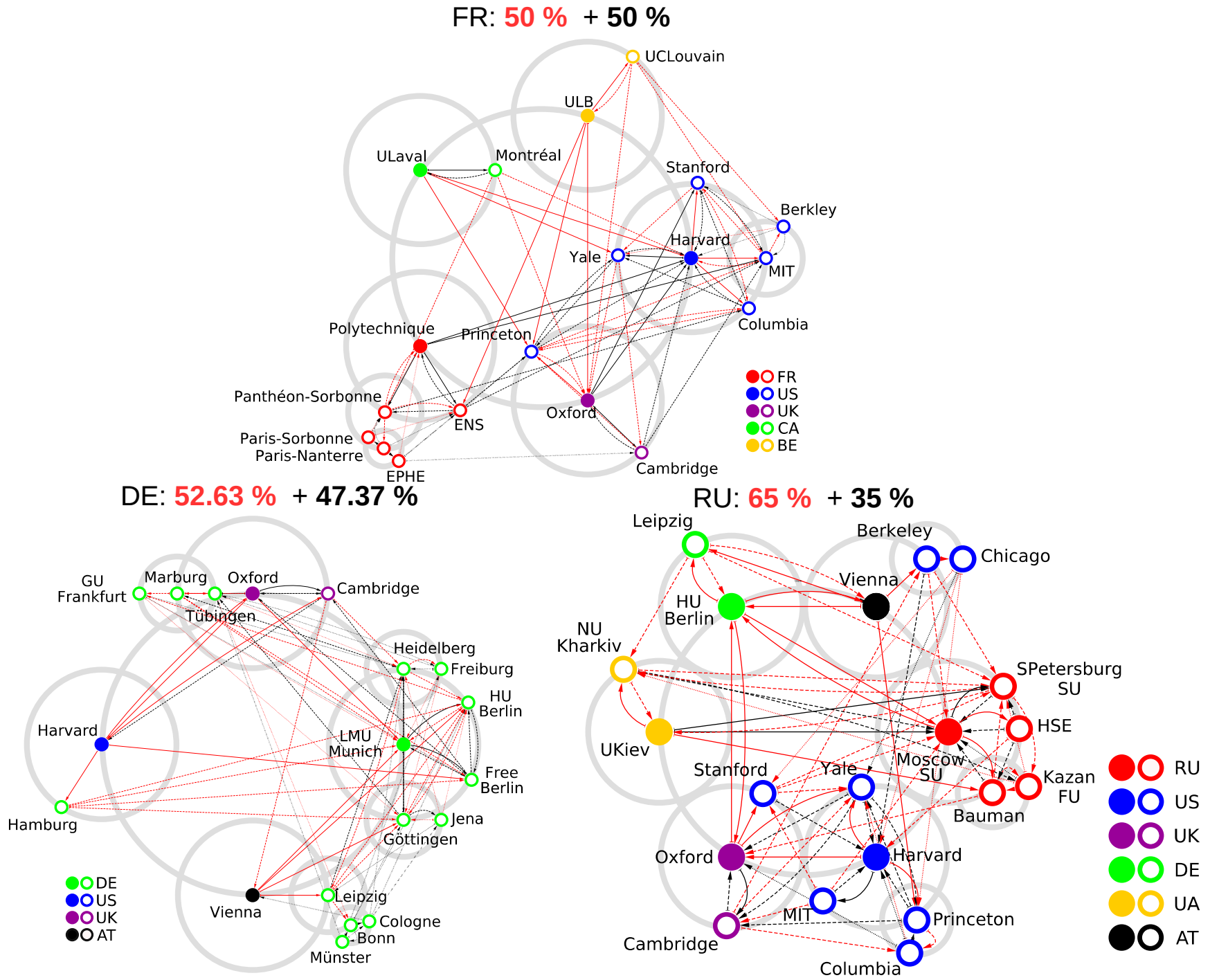


Fig. 5: Top 20 univ. reduced network for different editions: **Direct** and hidden links, solid lines (1st level), dashed lines (2nd), doted lines (3rd) and "\" symbol lines (4+)

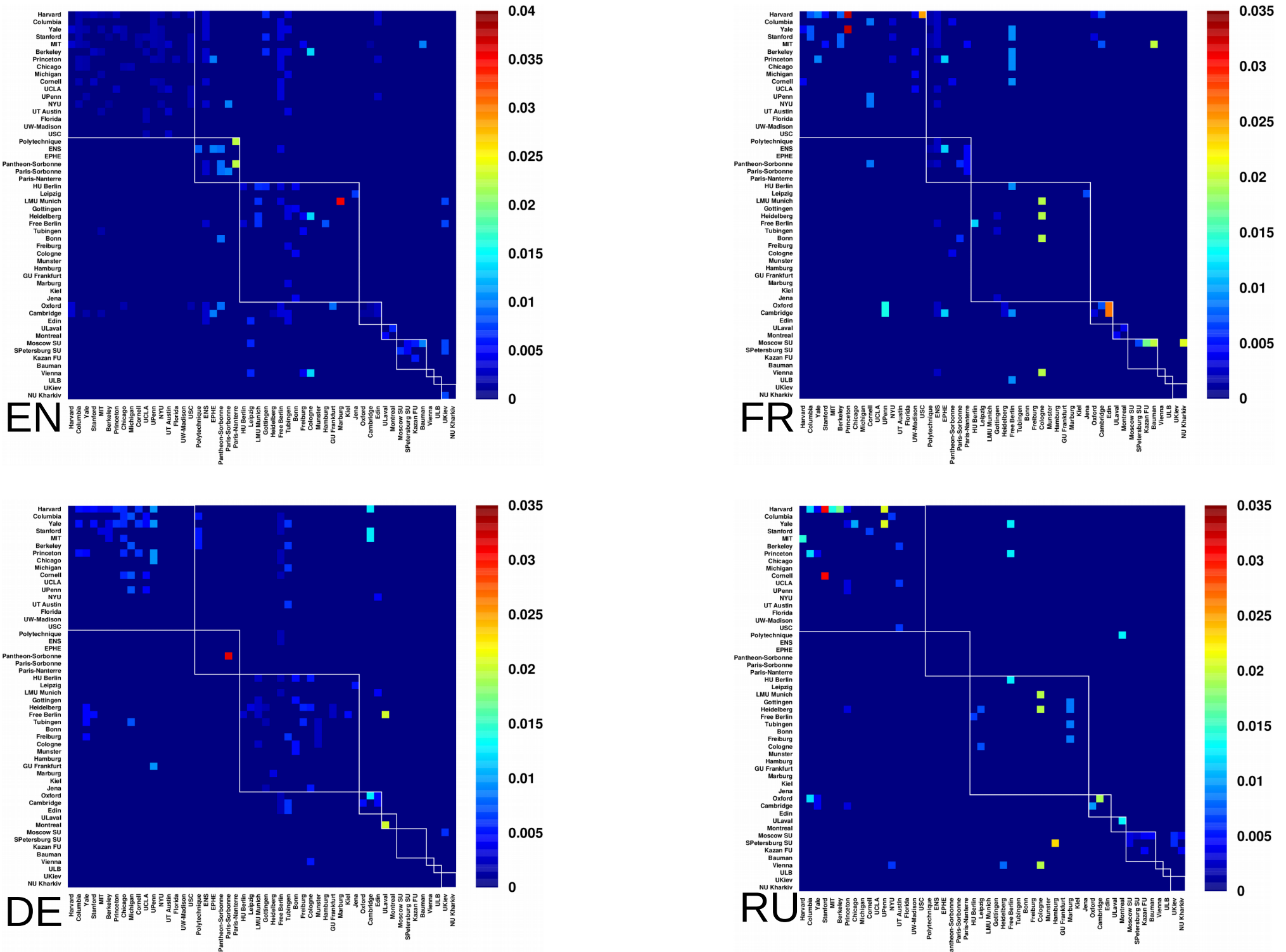


Fig. 7: $\mathbf{G_{rr}}$ for **EN**, **FR**, **DE** and **RU WIKI 2017** with a set of **52 universities** : 17 US; 6 FR ; 16 DE ; 3 UK ; 2 CA ; 4 RU ; 1 AU ; 1 BE and 2 UA.

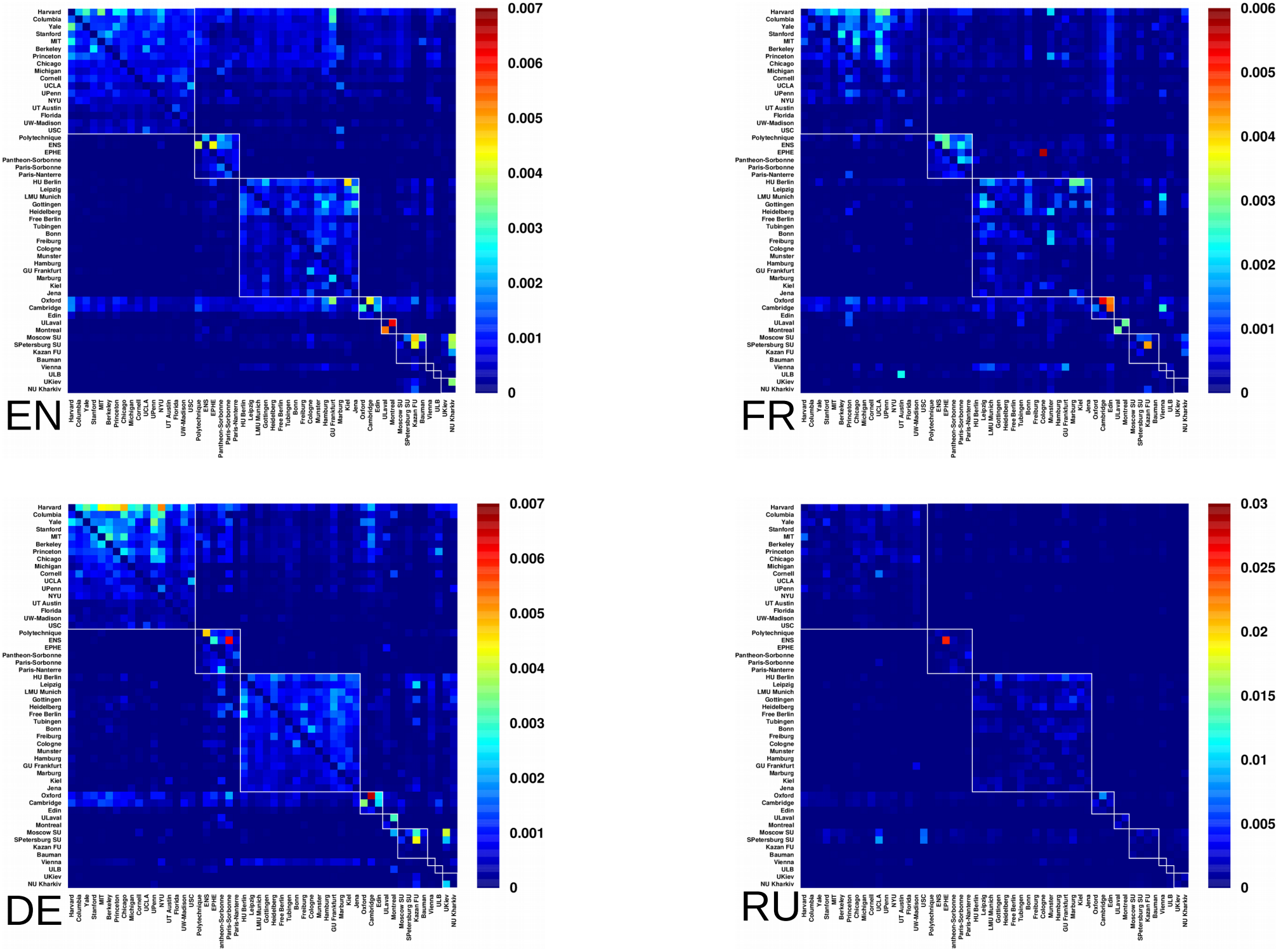


Fig. 7: $\mathbf{G_{qr}}$ for **EN**, **FR**, **DE** and **RU WIKI 2017** with a set of **52 universities** : 17 US; 6 FR ; 16 DE ; 3 UK ; 2 CA ; 4 RU ; 1 AU ; 1 BE and 2 UA.

7) Conclusion

- **Wikipedia** network gives us **numerous** interesting **application** in **different fields**
- **Google matrix** and **PageRank** algorithm + **University wikipedia** articles is an **efficient ranking** algorithm
- **REGOMAX** allows us to **focus** on tiny **communities** of interest keeping **all information** contained **in Wikipedia**
- **REGOMAX highlights** the **cultural differences** between **language editions** of **wikipedia**

7) Conclusion

- **Wikipedia network :**
  - Hidden relationship between political leaders (Frahm et al. 2016)
  - Terrorist groups (El Zant et al. 2018)
  - Application to biomedical topic such as infectious diseases (Rollin et al. 2019)
  - World influence and Interactions of Universities (Coquidé et al. 2019)
- **Trade network :**
  - Product with UN COMTRADE (Coquidé et al. 2019)
  - Economic activities with WTO (Coquidé et al. 2019)
  - Crypto-currency
- **Game network** :
  - Game of go (Coquidé et al. 2017)


Thank you