World influence from Wikipedia directed networks

Academic rankings, infectious diseases, cancers and drugs

José Lages
Célestin Coquidé, Guillaume Rollin
Equipe de Physique Théorique - Institut UTINAM
CNRS
Université de Bourgogne Franche-Comté

Dima Shepelyansky
Laboratoire de Physique Théorique
CNRS
Université Paul Sabatier, Toulouse

References:

J. L., A. Patt, D. L. Shepelyansky, The European Physical Journal B (2016) 89: 69
C. Coquidé, J. L., D. L. Shepelyansky, arXiv:1809.00332 (2018)
G. Rollin, J. L., D. L. Shepelyansky, bioRxiv (2018) doi:10.1101/424465

October 18th 2018

ApliGoogle project (2016-2018)
funded by MASTODONS CNRS Mission interdisciplinarité
Partners : LPT, CNRS, UPS, Toulouse
UTINAM, CNRS, UBFC, Besançon
I. Curie, Inserm, PSL, Paris
IRIT, CNRS, UPS, Toulouse

APEX project (2017-2020)
funded by Région Bourgogne Franche-Comté.

GNETWORKS project (2018-2021)
funded by ISITE-UBFC (PIA) ANR-15-IDEX-0003.

3 projects devoted to the physical analysis of complex networks
and the application of Google matrix based analysis to complex systems

Outline

Influence and interactions of ...

World Universities

Infectious diseases

Cancer and related drugs

... through the prism of Wikipedia

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"

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

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 

$+\beta$ 

$+\gamma$ 

$+\delta$ 

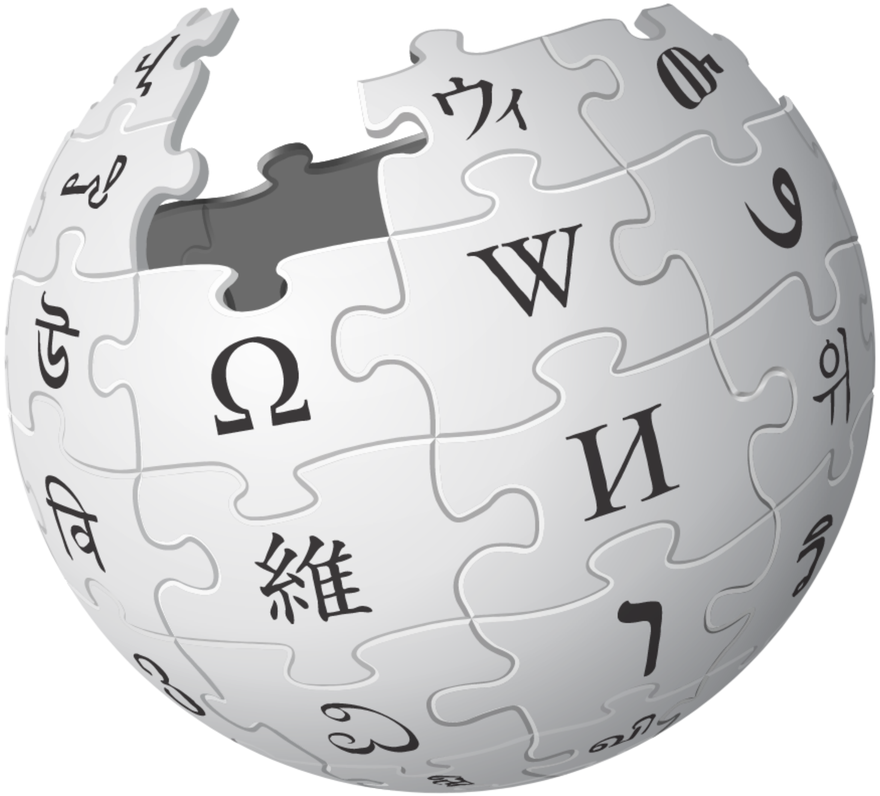
$+ \ldots$

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

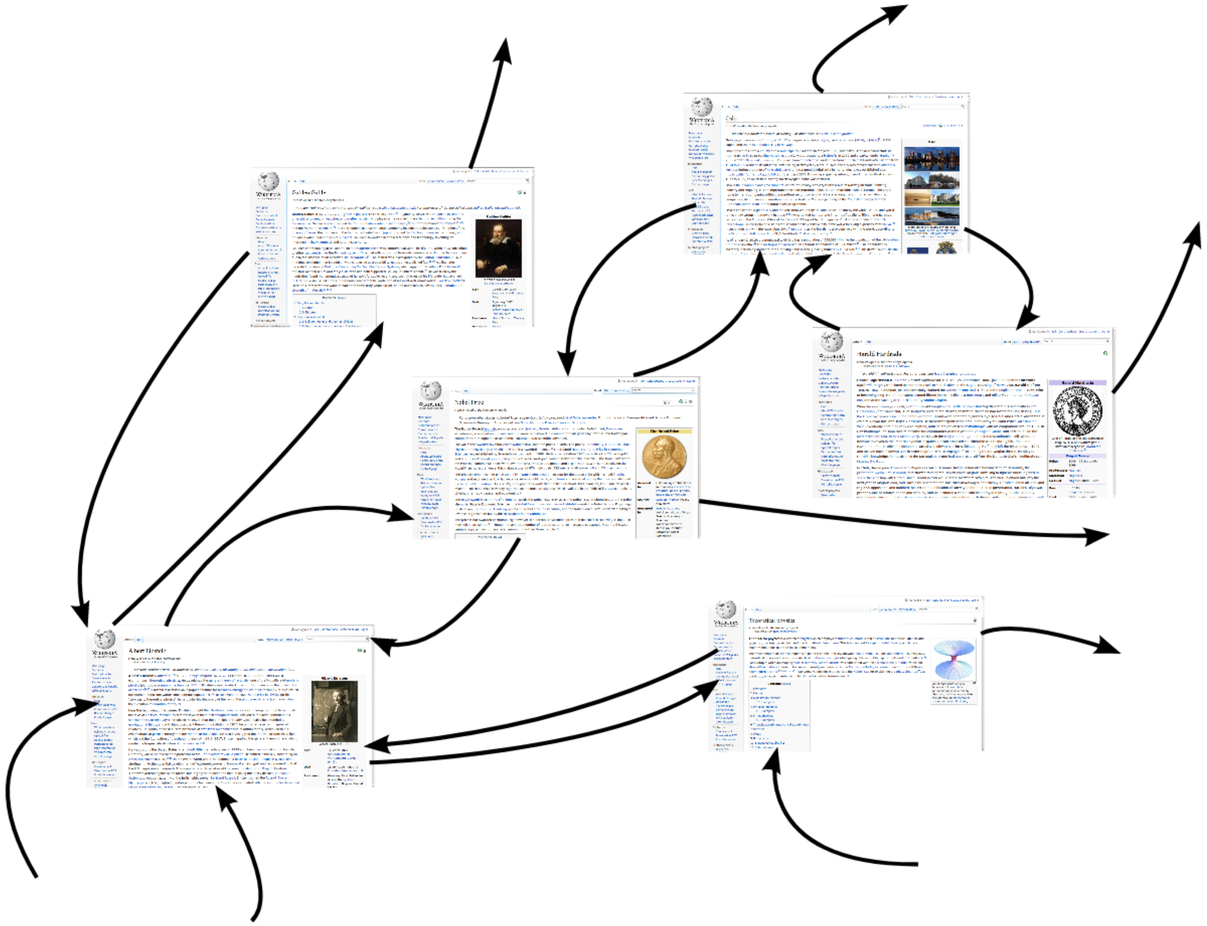
covering 59% of world population
and 68% total Wikipages

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



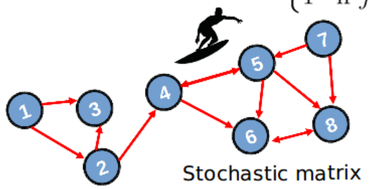
Google Matrix analysis

A Wikipedia language edition is characterized by N articles connected via N_ℓ hyperlinks.

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

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

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



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

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

Damping factor $\alpha = 0.85$

For a given wikiedition, we rank all the pages using:

PageRank algorithm
More a given page is pointed by important pages
more this page 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 random surfer ends at site 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 site 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)

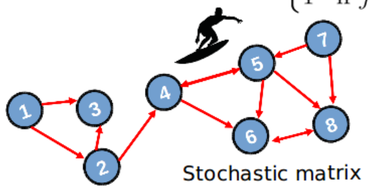
Google Matrix analysis

A Wikipedia language edition is characterized by N articles connected via N_ℓ hyperlinks.

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

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

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



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

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

Damping factor $\alpha = 0.85$

For a given wikiedition, we rank all the pages using:

PageRank algorithm
More a given page is pointed by important pages
more this page 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 random surfer ends at site 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 site 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)

Stochastic matrix $S^*_{ij} = \begin{cases} A_{ji}/k_{in}(j) & \text{if } k_{in}(j) \neq 0 \\ 1/N & \text{otherwise} \end{cases}$

Google matrix $G^*_{ij} = \alpha S^*_{ij} + (1-\alpha)/N$

Damping factor $\alpha = 0.85$

For a given wikiedition, we rank all the pages using:

CheiRank algorithm
More a given page points to important pages
more this page is important
(Measure of communicativity)

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

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

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

After a sufficiently long journey, P^*_i is the probability that a random surfer started his journey from site i

The CheiRank index is $K^* \in \{1, \ldots, N\}$
$K^* = 1$ for page with highest P^*_i
$K^* = N$ for page with lowest P^*_i

(Chepelianskii '10, Ermann *et al.* '12, Zhirov *et al.* '10)

For a given wikiedition, we rank all the pages using:

2DRank algorithm
(Measure influence and communicativity)

The 2DRank index is $K_2 \in \{1, \ldots, N\}$
Assuming a page has a PageRank index K and a CheiRank index K^*
the 2DRank index is $K_2 = \max\{K, K^*\}$

(Zhirov *et al.* '10)

Wikipedia Ranking of World Universities

From the Page/Chei/2DRanking of each of the 24 Wikipedia language editions, we extract the rank index of pages devoted to "Universities" and we establish top 100 for each editions and for each algorithms.

Example: PageRank algorithm applied to frwiki (March '13) gives the following top 3:

1- K=904 "Université de Harvard"
2- K=1549 "Ecole Polythechnique"
3- K=1558 "Université d'Oxford"

Let $R_{U,E,A}$ be the rank index of an university U appearing in the top 100 of the Wikipedia edition E obtained via algorithm A

The global ranking is obtained defining the ranking score $\Theta_{U,A} = \sum_E (101 - R_{U,E,A})$

Results: 1024 ranked universites with PageRank algorithm, 1378 with CheiRank algorithm and 1559 with 2DRank algorithm

Universities from 142 countries.

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

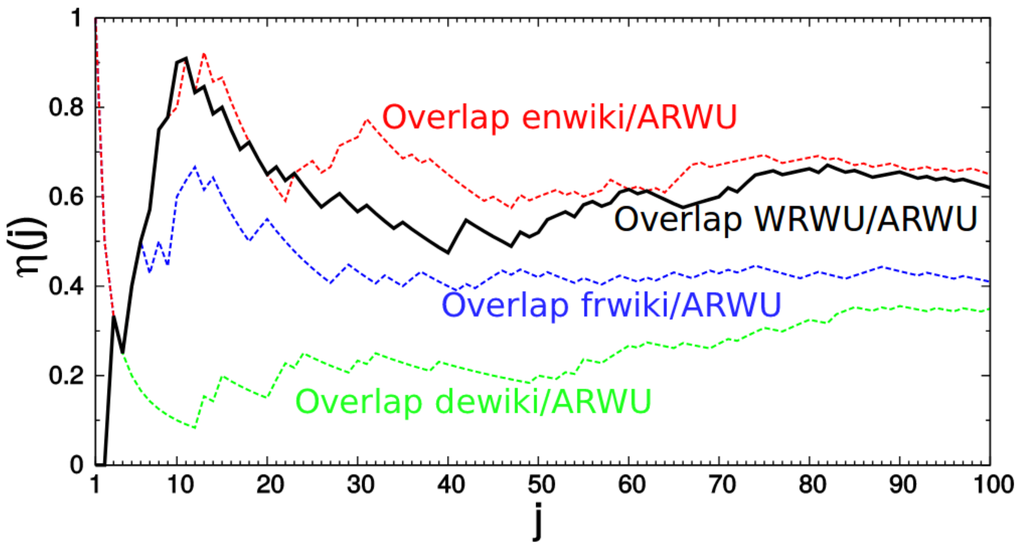
60% overlap
between top 100s
WRWU and ARWU

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

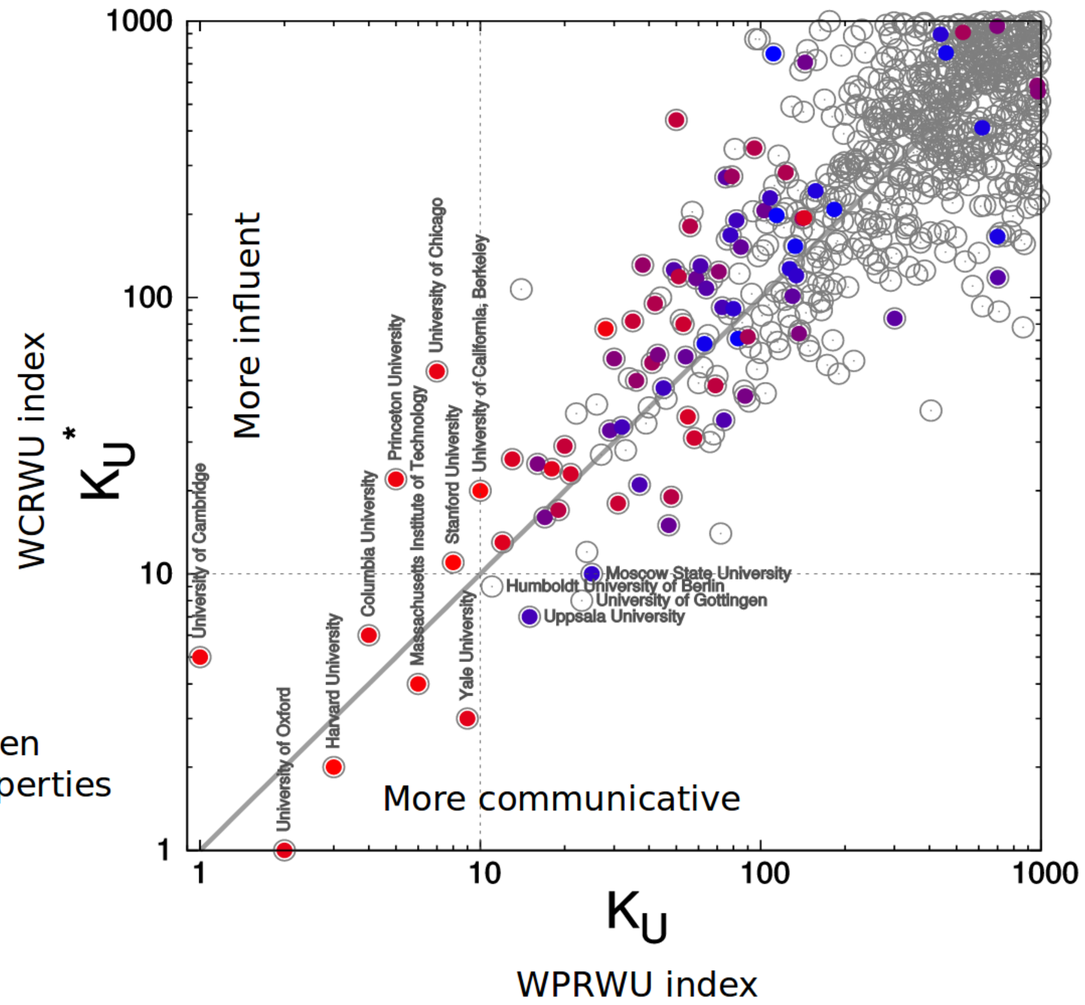
Definitively, as ARWU, WRWU measures academic excellence, but not only ...

Rank	WPRWU	Θ_{PR}	N_a
1st	University of Cambridge	2272	24
2nd	University of Oxford	2247	24
3rd	Harvard University	2112	22
4th	Columbia University	2025	23
5th	Princeton University	1887	23
6th	Massachusetts Institute of Technology	1869	21
7th	University of Chicago	1783	22
8th	Stanford University	1765	21
9th	Yale University	1716	20
10th	University of California, Berkeley	1557	19

Rank	WCRWU	Θ_{CR}	N_a
1st	University of Oxford	1191	18
2nd	Harvard University	1025	17
3rd	Yale University	1021	16
4th	Massachusetts Institute of Technology	816	16
5th	University of Cambridge	803	11
6th	Columbia University	779	14
7th	Uppsala University	751	11
8th	University of Göttingen	735	13
9th	Humboldt University of Berlin	703	12
10th	Moscow State University	699	14

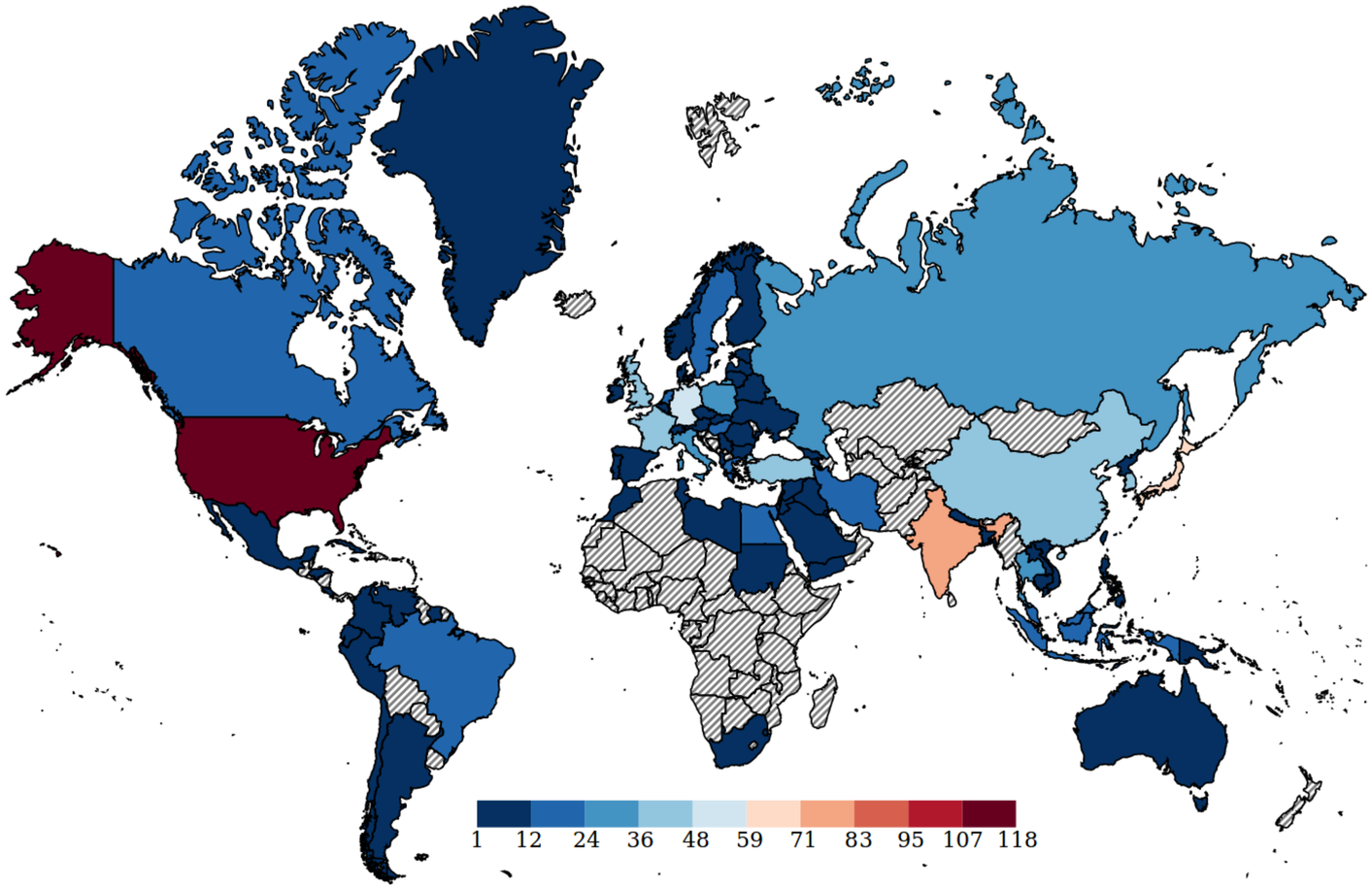
Rank	W2RWU	Θ_{2R}	N_a
1st	University of Cambridge	942	16
2nd	Columbia University	786	11
3rd	Stanford University	712	11
4th	Harvard University	683	11
5th	Yale University	609	11
6th	Princeton University	596	11
7th	Massachusetts Institute of Technology	581	10
8th	Humboldt University of Berlin	578	10
9th	Nanjing University	516	8
10th	Johns Hopkins University	511	9

Rank	ARWU	WPRWU	WCRWU	W2RWU
1st	Harvard University	−2	−1	−3
2nd	Stanford University	−6	−9	−1
3rd	University of California, Berkeley	−7	−17	−13
4th	Massachusetts Institute of Technology	−2	0	−3
5th	University of Cambridge	+4	0	+4
6th	California Institute of Technology	−22	−71	−124
7th	Princeton University	+2	−15	+1
8th	Columbia University	+4	+2	+6
9th	University of Chicago	+2	−45	−70
10th	University of Oxford	+8	+9	−2



Approximative balance between influence and communicative properties of universities in Wikipedia

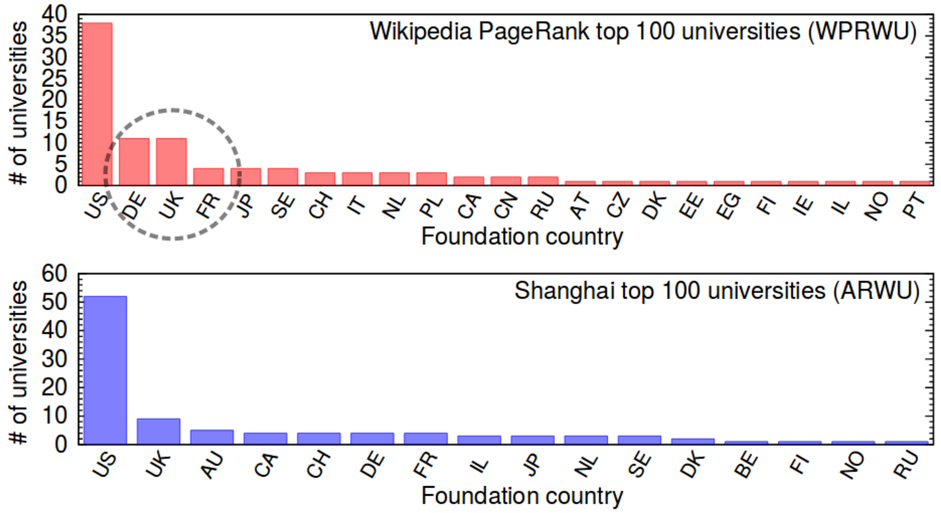
Geographical distribution of universities in WRWU



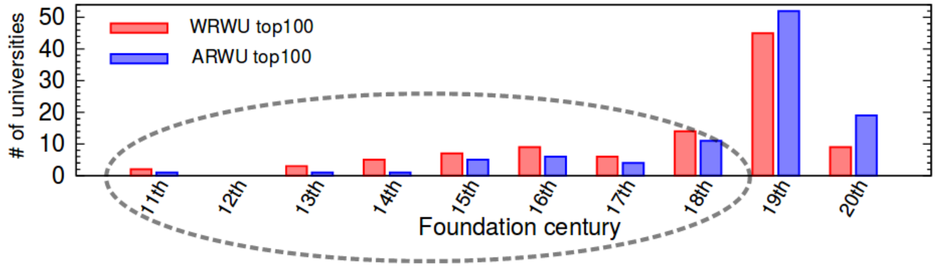
As in the other rankings such as ARWU,
US universities still dominate,

BUT ...

... **less US universities**
and **more european universities** among top 100, ...



... also **more older universities** among top 100.



"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

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

Universities in Top 43 were founded before XXth 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

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

71 19 Waseda University

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

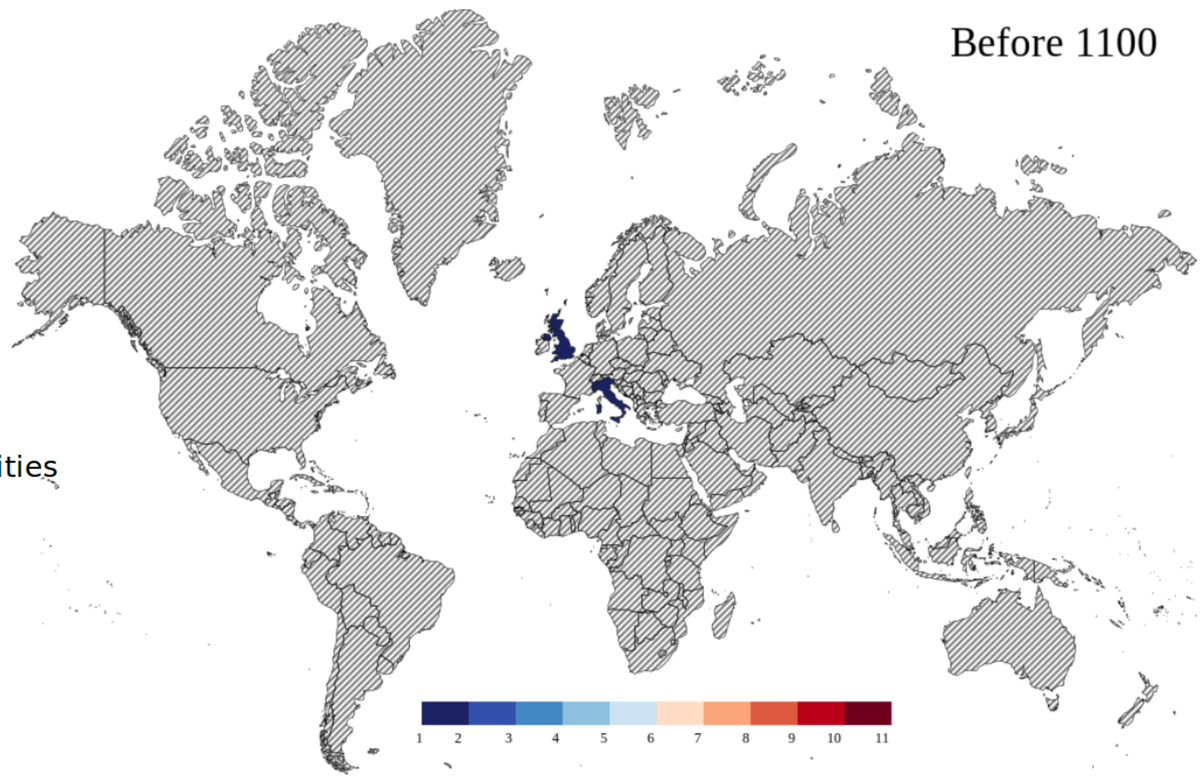
Universities in Top 43 were founded before XXth century

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	Ecole Normale Supérieure	FR	FR	18
38	570	14	University of Manchester	UK	EN	19
39	556	13	Ecole 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	13	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

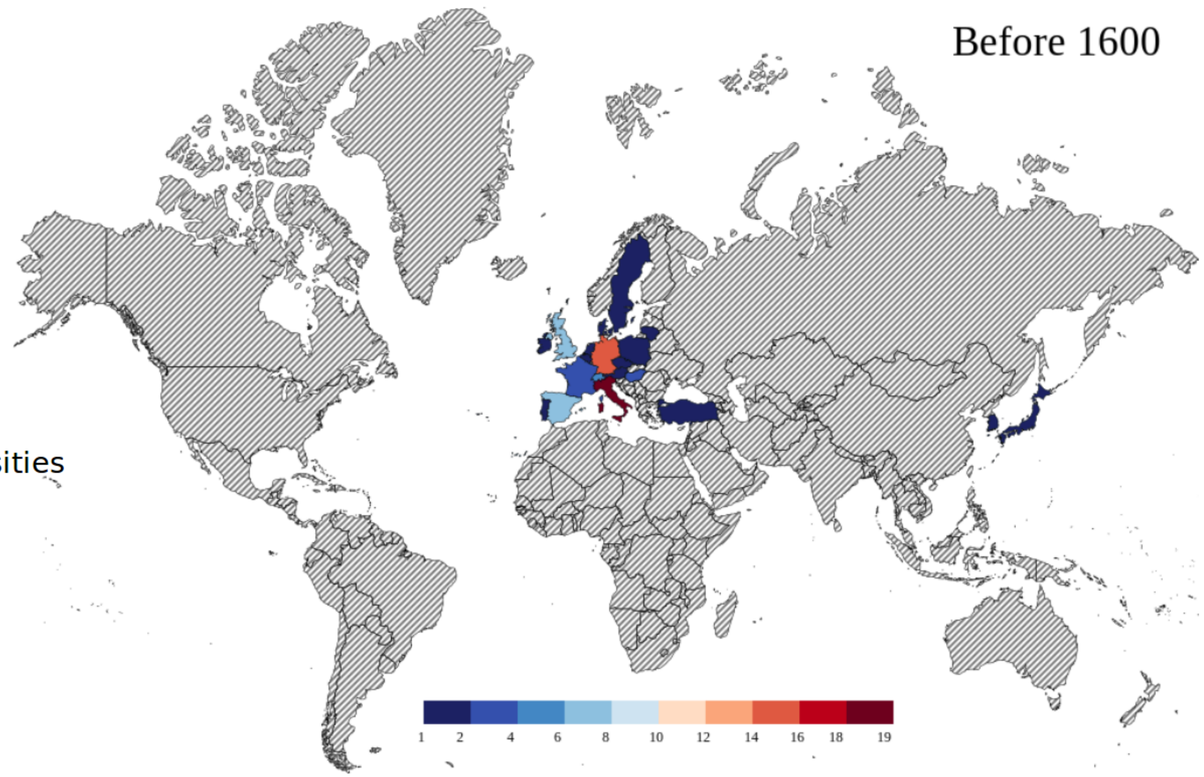
Quite rigid club
of first universities
"not willing" to accept
new members

Timeline
Evolution through centuries

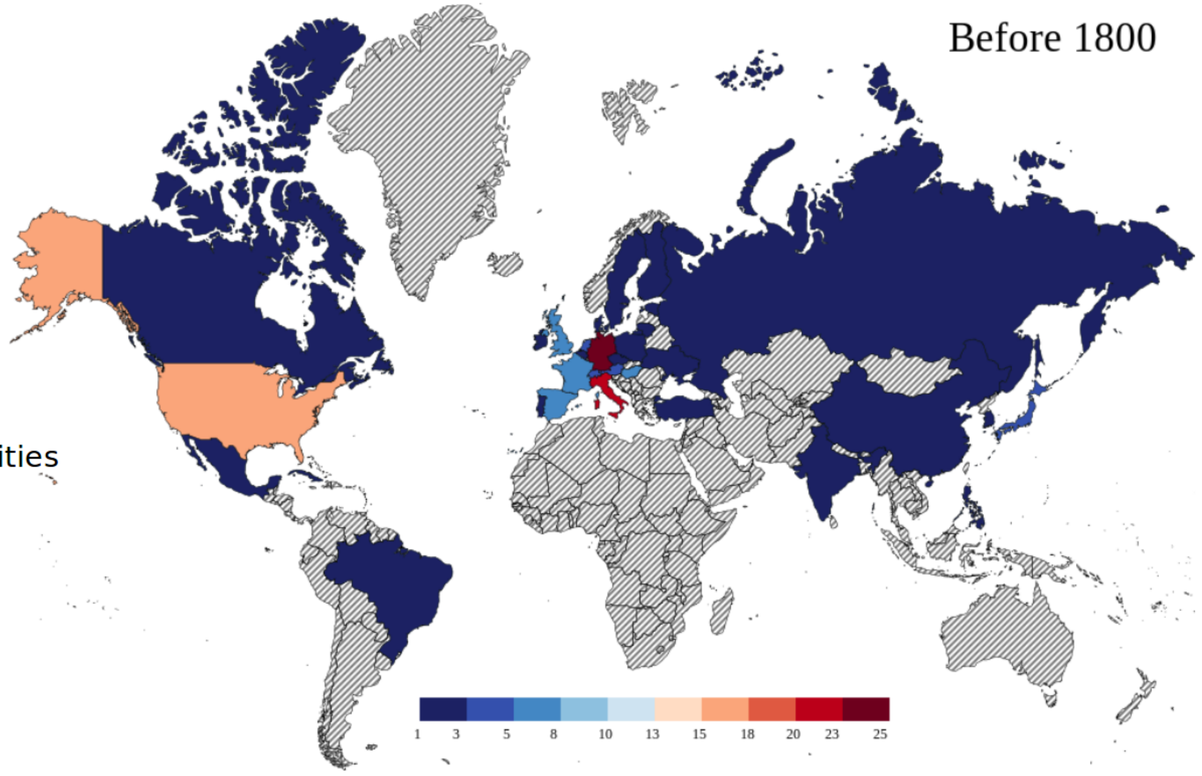
First universities



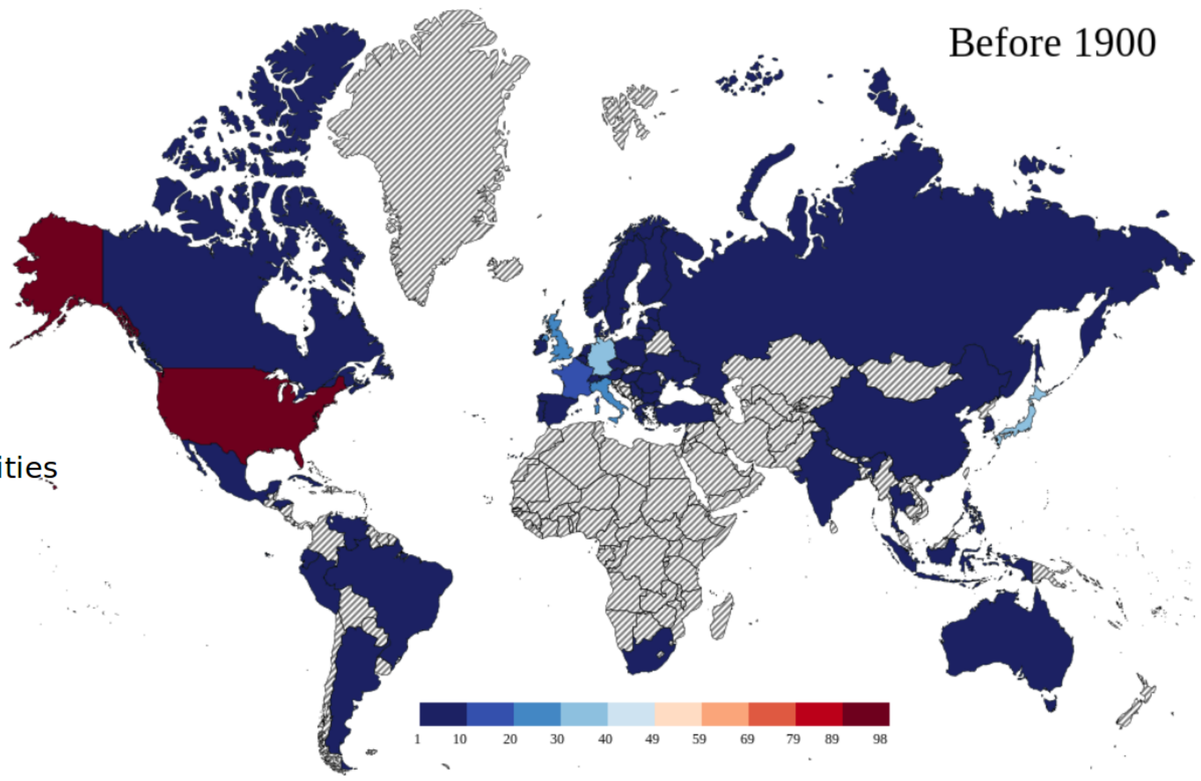
Dominance of European universities



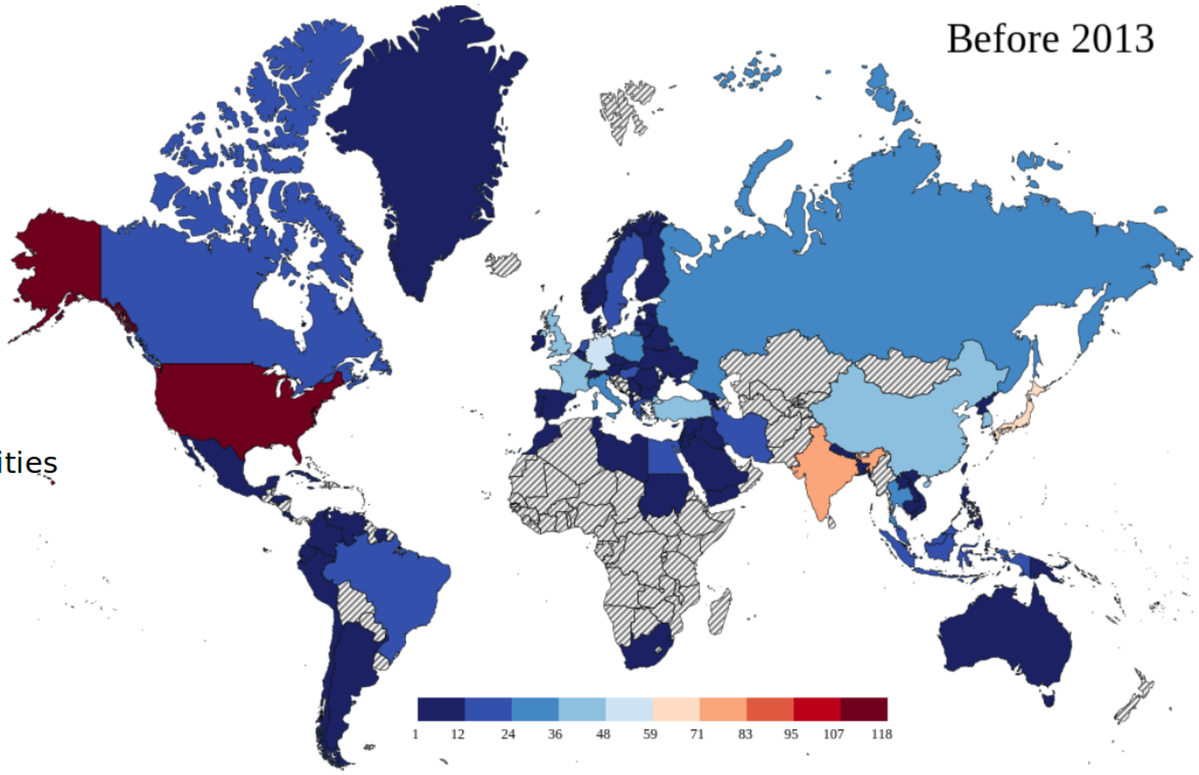
Emergence of US universities



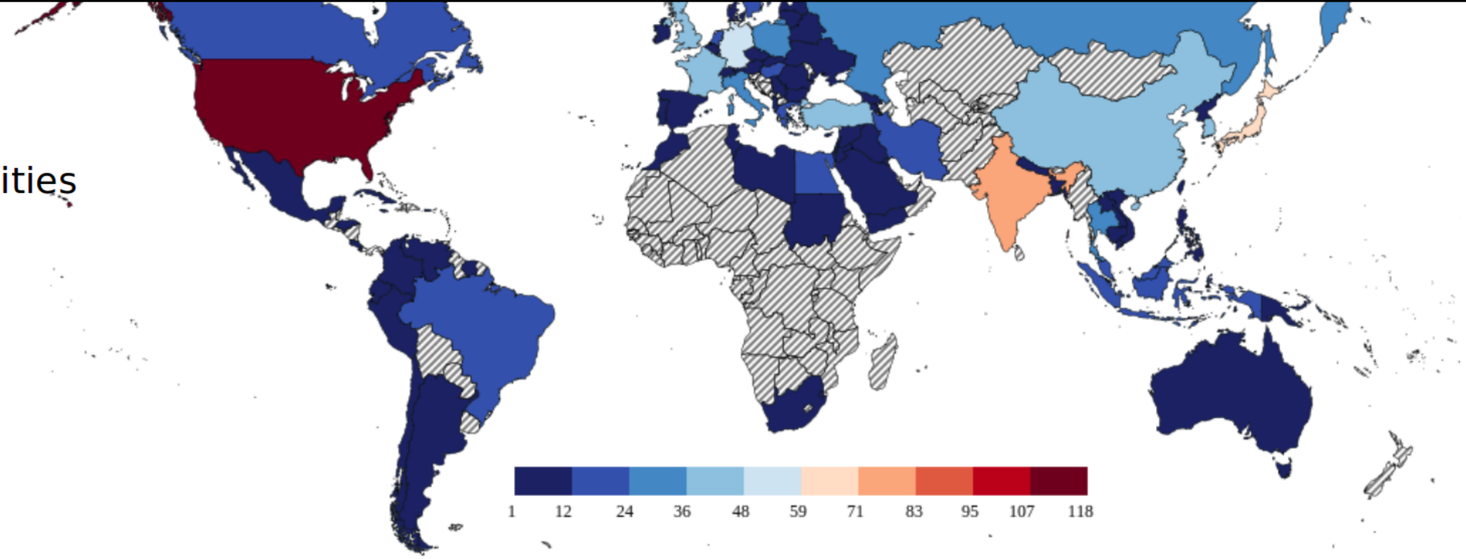
Dominance of US universities



Emergence of Asian universities



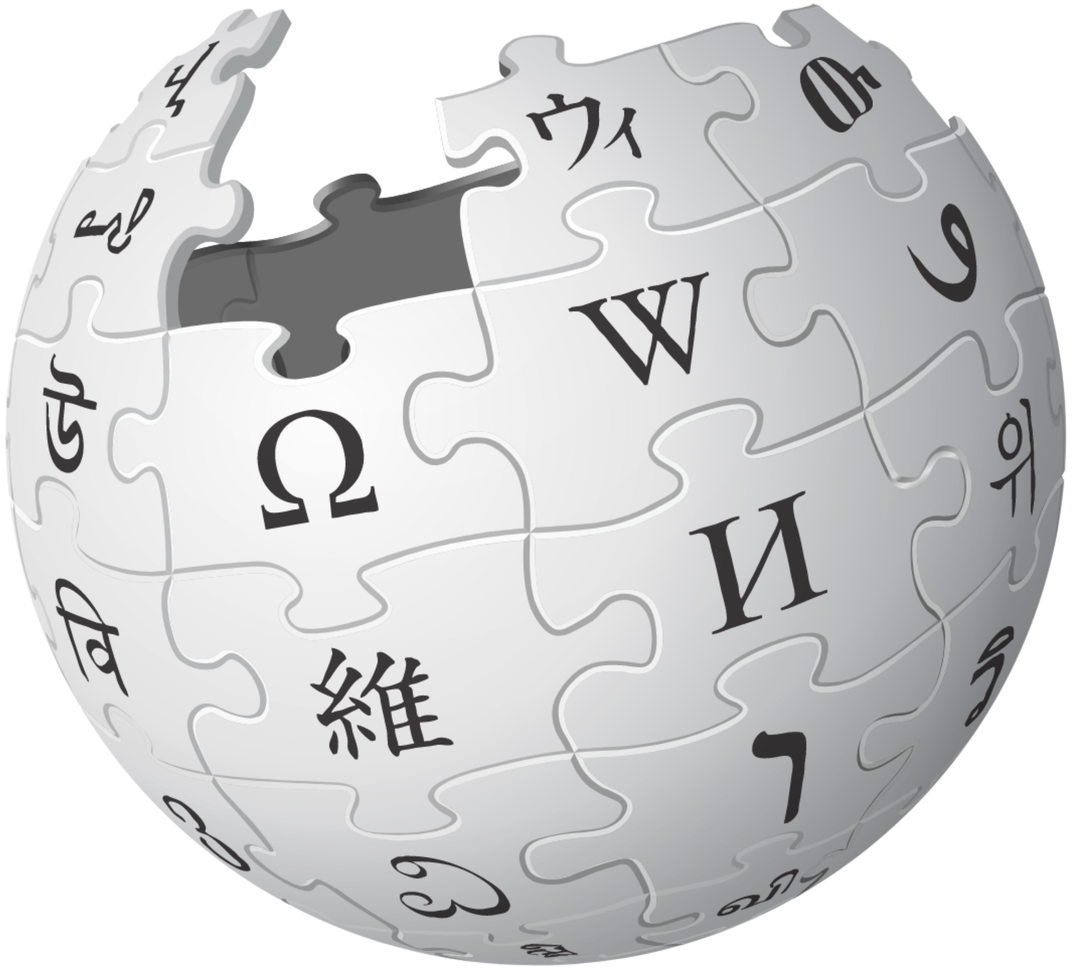
Emergence of Asian universities



What's next ? dominance of Asian universities ?

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.

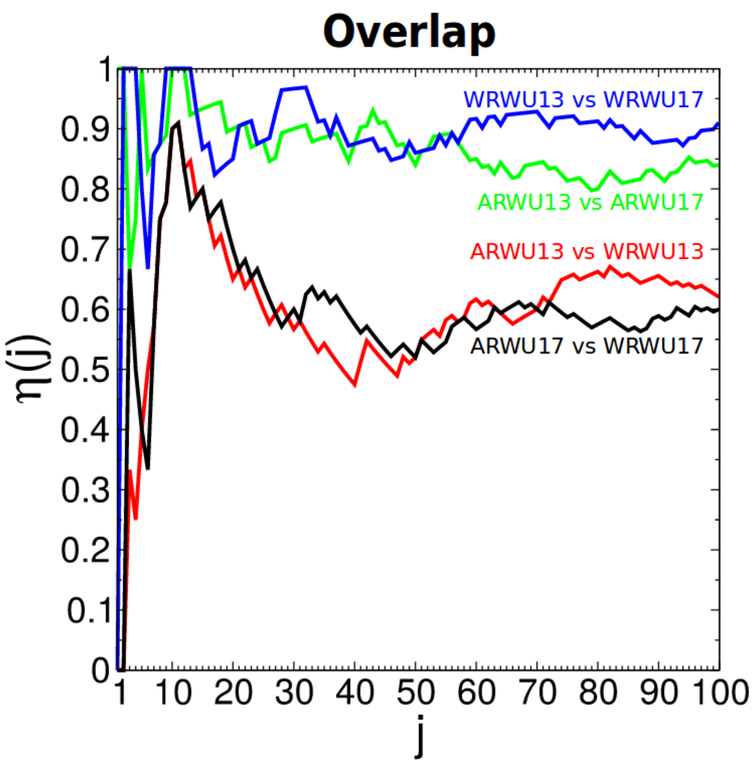
Universal ranking ?

Want to check the position of your Alma Mater ?

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

http://perso.utinam.cnrs.fr/~lages/datasets/WRWU17/

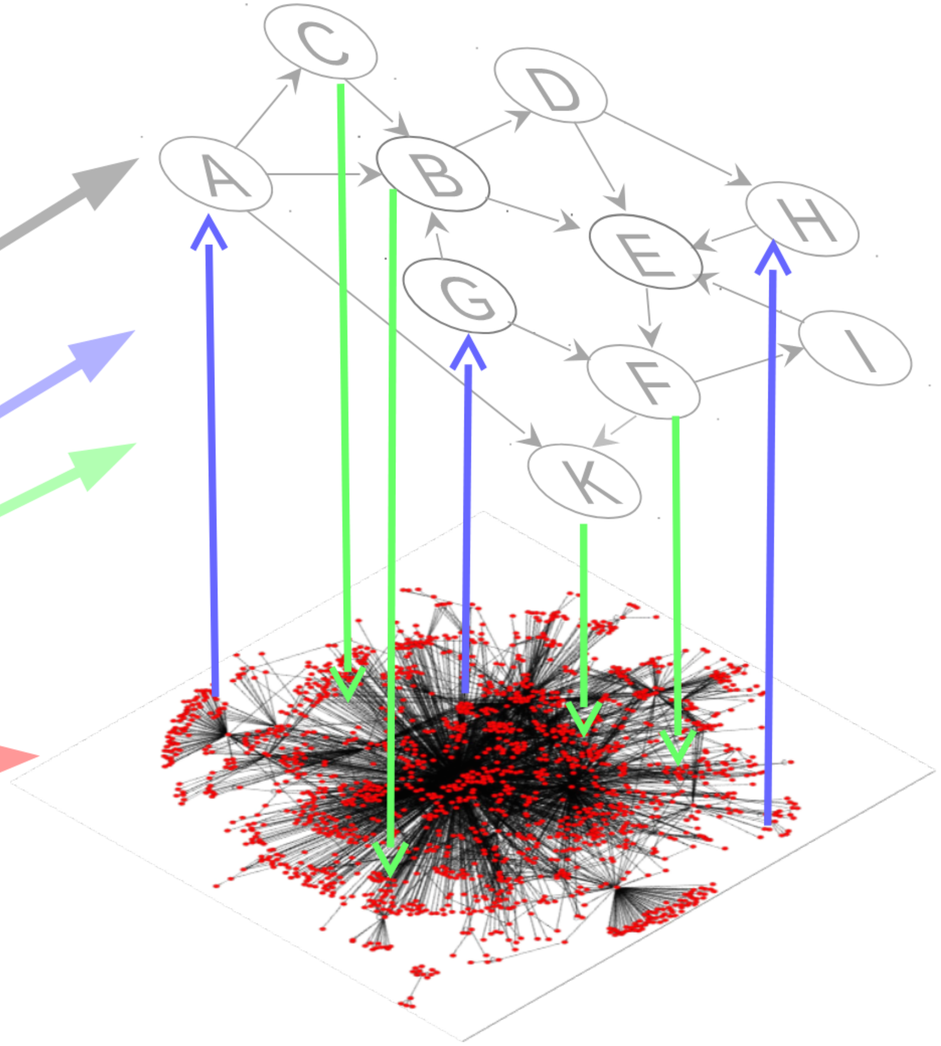
Wikipedia PageRanking of World Universities WRWU 2013		Wikipedia PageRanking of World Universities WRWU 2017
University of Cambridge	1st	University of Oxford
University of Oxford	2nd	University of Cambridge
Harvard University	3rd	Harvard University
Columbia University	4th	Columbia University
Princeton University	5th	Yale University
MIT	6th	University of Chicago
University of Chicago	7th	Princeton University
Stanford University	8th	Stanford University
Yale University	9th	MIT
University of California, Berkeley	10th	University of California, Berkeley



Robustness of WRWU

Reduced Google matrix

Consider a network with $N \gg 1$ nodes.
Consider a sub-network of $N_r \ll N$ nodes of interest.
The Google matrix of the whole network and the associated PageRank vector can be written as

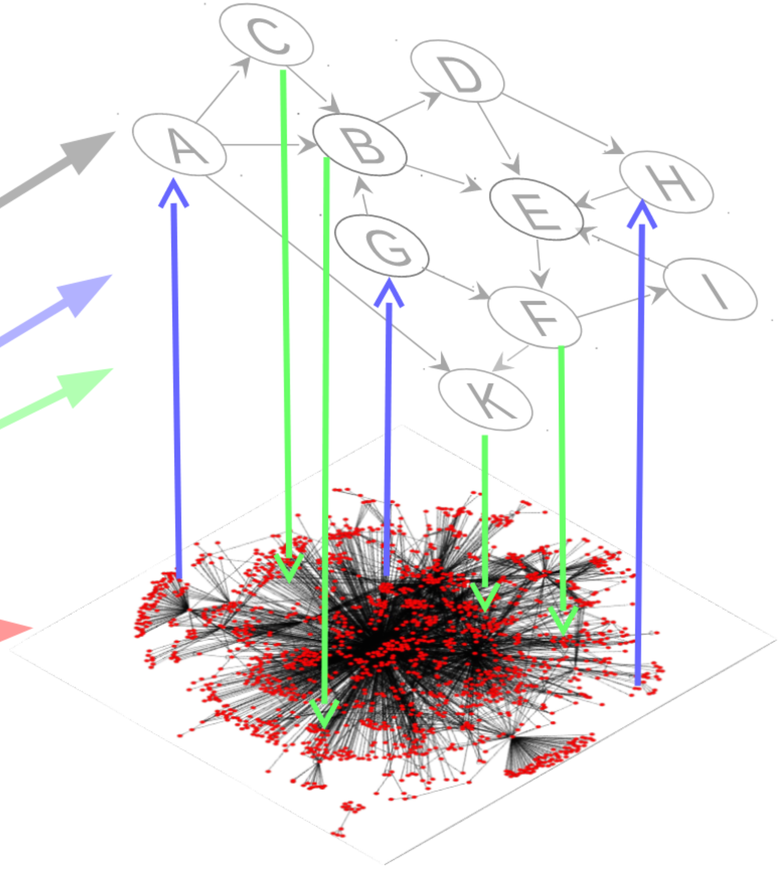


$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_{rr} & \mathbf{G}_{rs} \\ \mathbf{G}_{sr} & \mathbf{G}_{ss} \end{pmatrix},$$

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_r \\ \mathbf{P}_s \end{pmatrix} \qquad \mathbf{GP} = \mathbf{P}$$

Reduced Google matrix

Consider a network with $N \gg 1$ nodes.
Consider a sub-network of $N_r \ll N$ nodes of interest.
The Google matrix of the whole network and the associated PageRank vector can be written as



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

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_r \\ \mathbf{P}_s \end{pmatrix} \qquad \mathbf{G}\mathbf{P} = \mathbf{P}$$

We define, $\mathbf{G}_R$, the reduced Google matrix associated to the sub-network of size N_r such as

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

The reduced Google matrix can be written

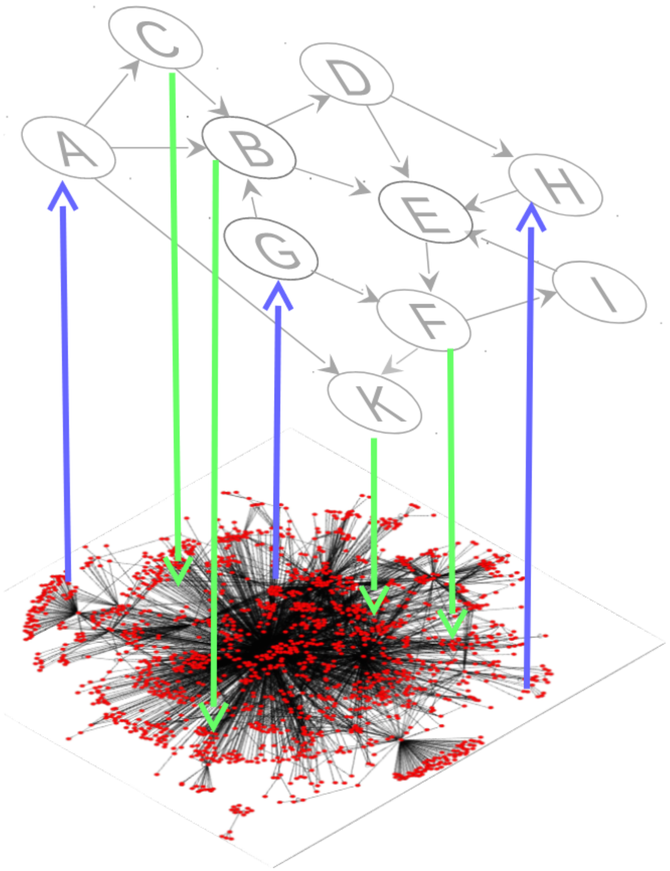
$$\mathbf{G}_R = \mathbf{G}_{rr} + \mathbf{G}_{rs}\left(\mathbf{1} - \mathbf{G}_{ss}\right)^{-1}\mathbf{G}_{sr}$$

Contribution from direct links

Contribution from indirect links (scattering term)

Very slow convergence since the leading eigenvalue λ_c ofi $\mathbf{G}_{ss} \sim \mathbf{G}$ s very close to 1

K. M. Frahm, and D. L. Shepelyansky, arXiv:1602.02394 [physics.soc-ph]
J . Lages, D. Shepelyansky, A. Zinovyev, PLoS ONE 13(1): e0190812 (2018)



Reduced Google matrix

Consider a network with $N \gg 1$ nodes.
Consider a sub-network of $N_r \ll N$ nodes of interest.

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

$$(\mathbf{1} - \mathbf{G}_{ss})^{-1} = \mathbf{P}_c (1 - \lambda_c)^{-1} + \mathbf{Q}_c \sum_{l=0}^{\infty} (\mathbf{Q}_c \mathbf{G}_{ss} \mathbf{Q}_c)^{l}$$

Contribution from "PageRank"

$$\mathbf{P}_c = \psi_R \psi_L^T \text{ with } \begin{cases} G_{ss}\psi_R = \lambda_c \psi_R \\ \psi_L^T G_{ss} = \lambda_c \psi_L \end{cases}$$

$$\mathbf{Q}_c = \mathbf{1} - \mathbf{P}_c$$

The reduced Google matrix can be written

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

Contribution from direct links

Contribution from indirect links (scattering term)

Very slow convergence since the leading eigenvalue λ_c ofi $\mathbf{G}_{ss} \sim \mathbf{G}$ s very close to 1

The reduced Google matrix can be rewritten

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

Contribution from direct links

Contribution from hidden links

Contribution from "PageRank"

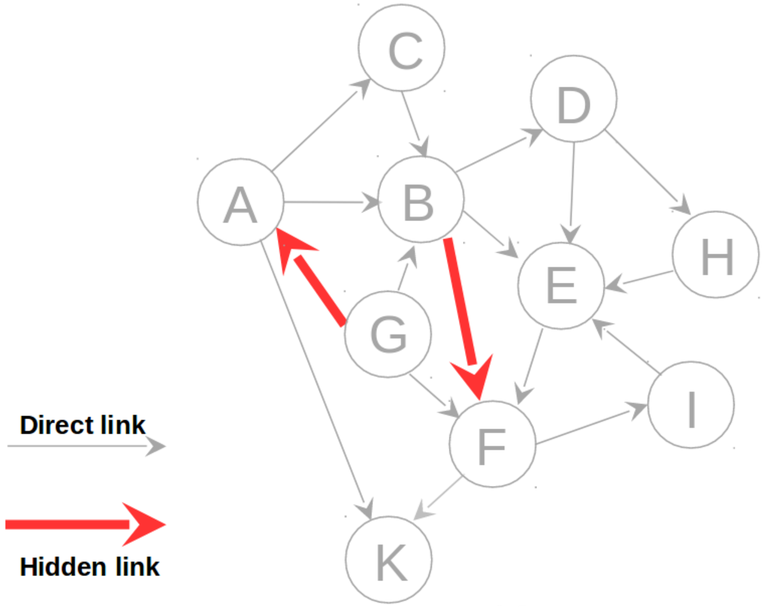
The reduced Google matrix can be rewritten

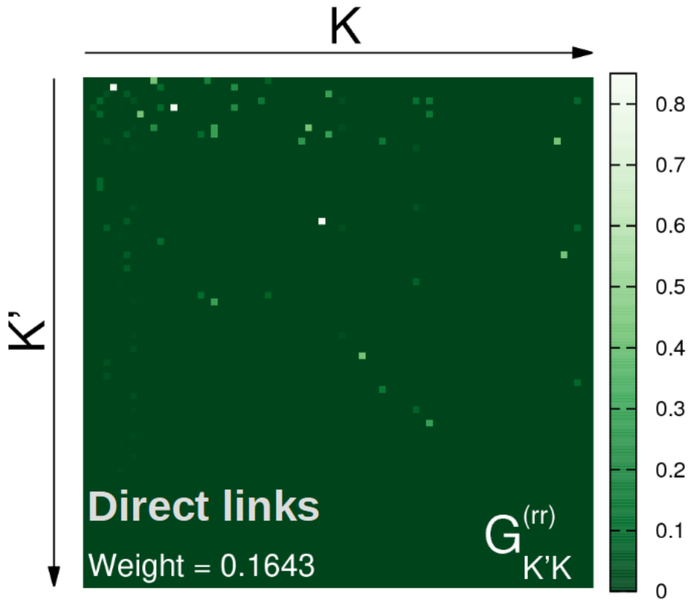
$$\mathbf{G}_R = \mathbf{G}_{rr} + \mathbf{G}_{pr} + \mathbf{G}_{qr}$$

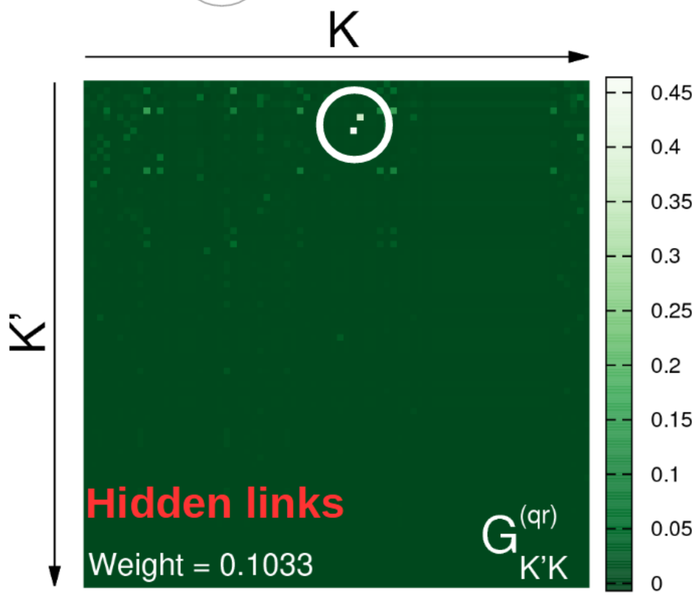
Contribution from direct links

Contribution from hidden links

Contribution from "PageRank"

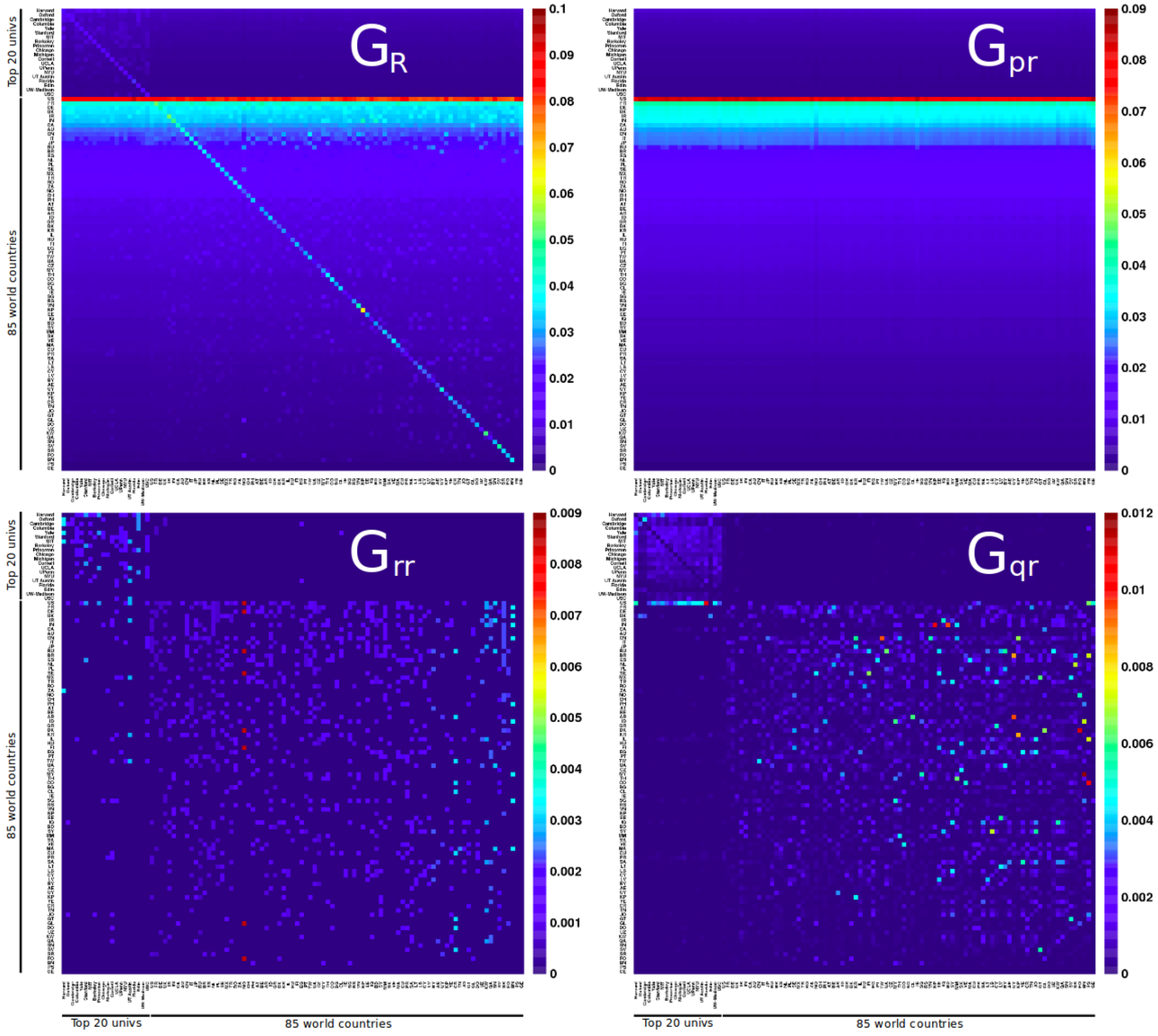






2017 English Wikipedia

Reduced Google matrix for the subset of interest
"top 20 universities" + "world countries"



PageRank sensitivity

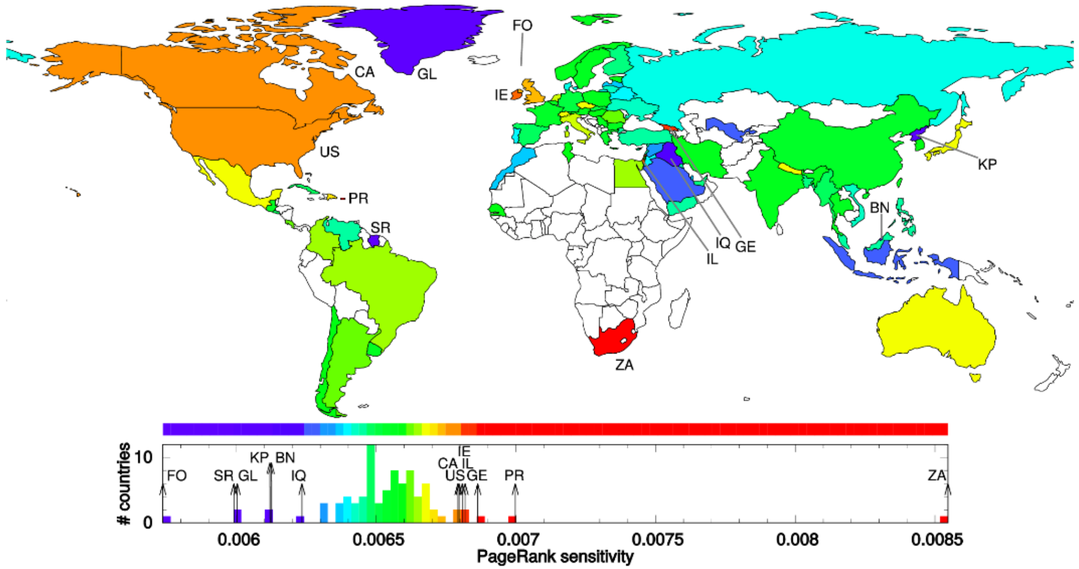
(to the change of a reduced Google matrix link)

$$D(j \to i, k) = \frac{1}{P_k(0)} \lim_{\delta_{ij} \to 0} \left(\frac{P_k(\delta_{ij}) - P_k(0)}{\delta_{ij}} \right)$$

δ_{ij} is the infinitesimal change
in $j \to i$ reduced Google matrix link
$P_k(\delta_{ij})$ is the kth component of the associated PageRank vector

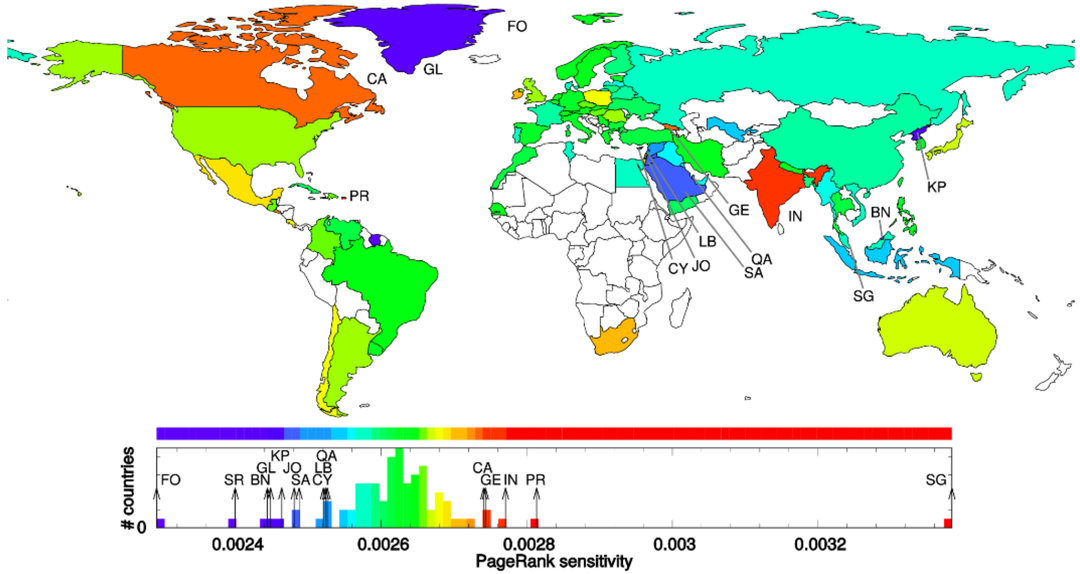
Harvard University

$D(\text{Harvard} \rightarrow \text{Country}, \text{Country})$



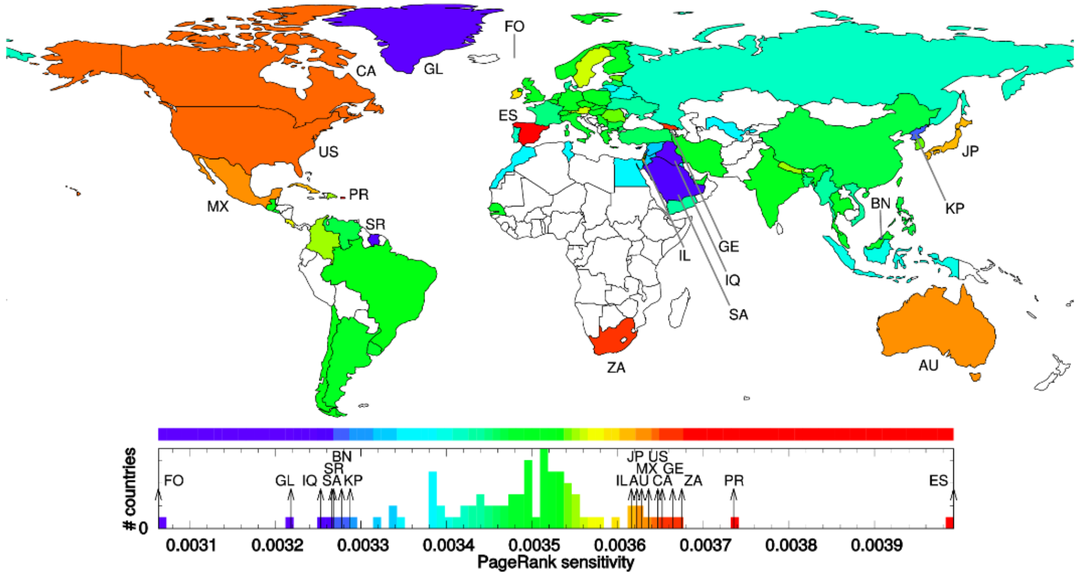
University of Chicago

$D(\text{Chicago} \rightarrow \text{Country}, \text{Country})$



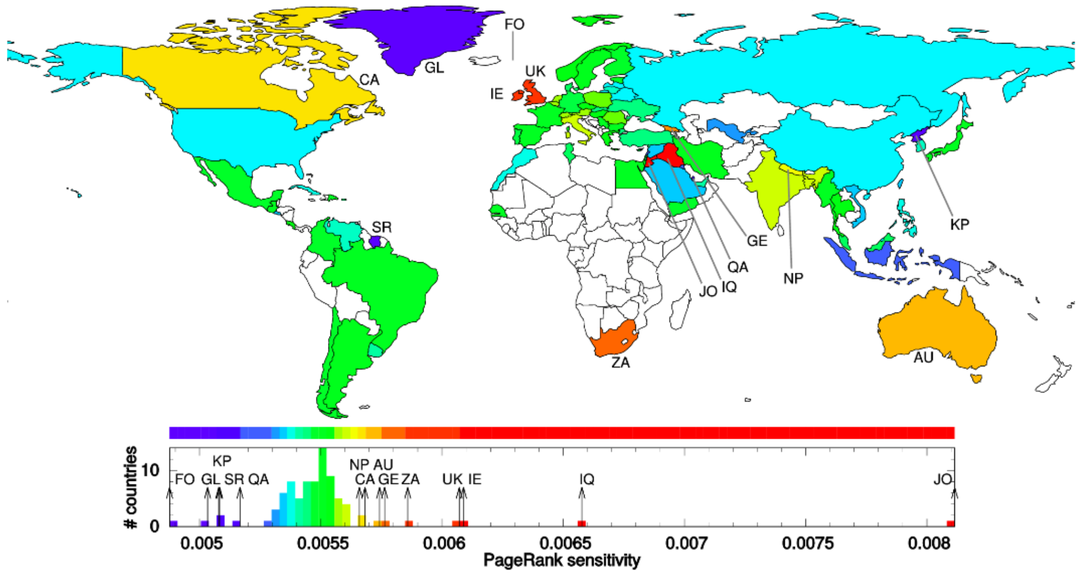
Stanford University

$D(\text{Stanford} \rightarrow \text{Country}, \text{Country})$



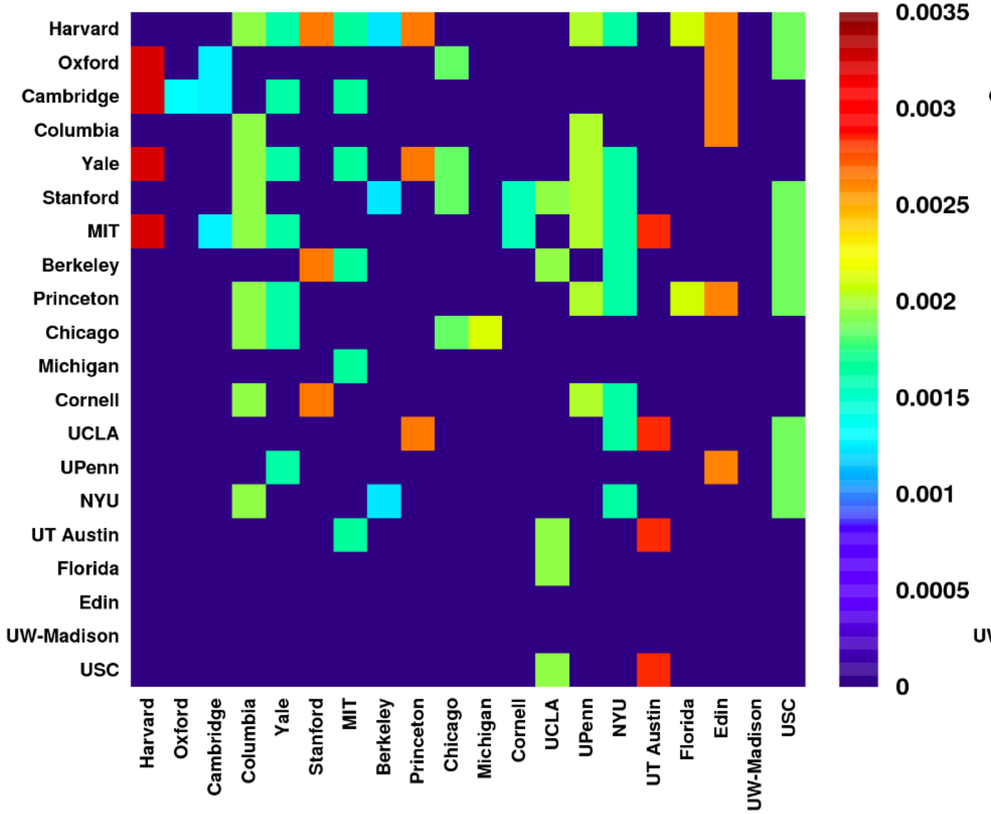
University of Oxford

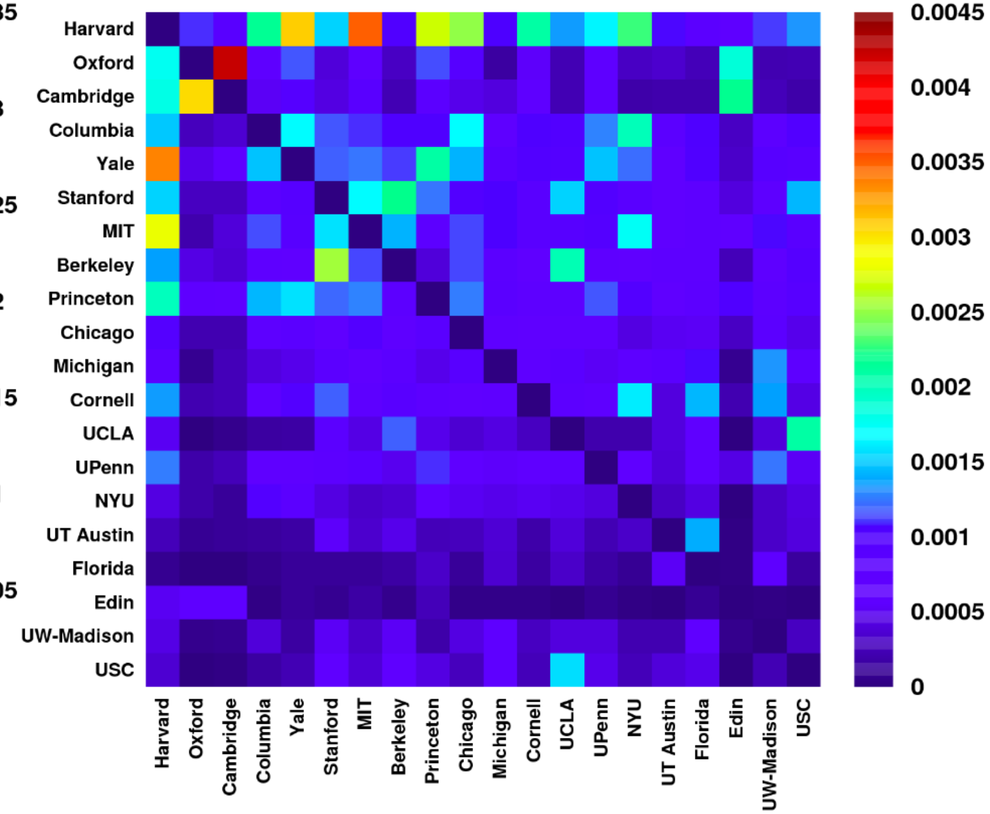
$D(\text{Oxford} \rightarrow \text{Country}, \text{Country})$



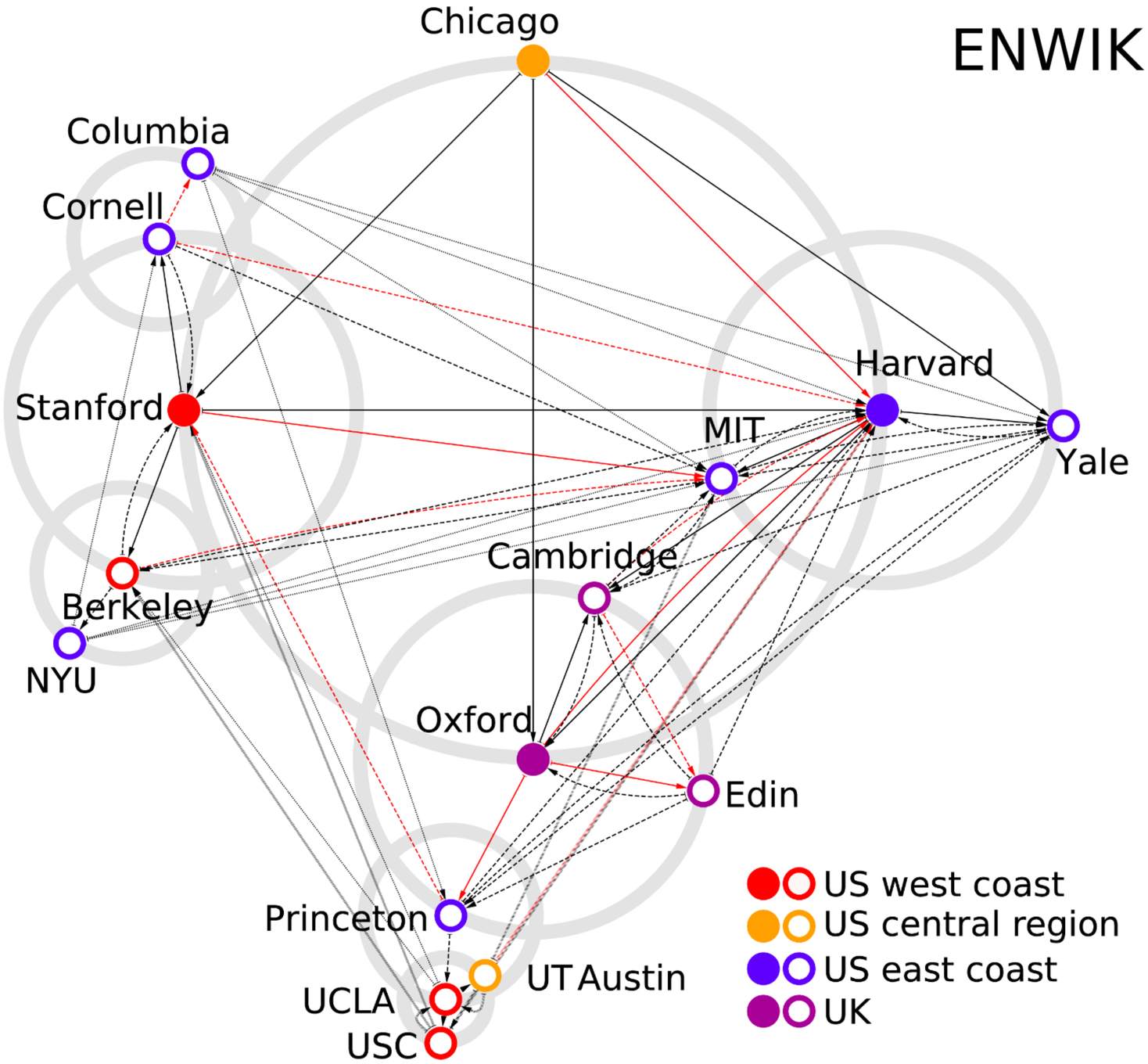
2017 English Wikipedia

Reduced Google matrix for the subset of interest
"top 20 universities"

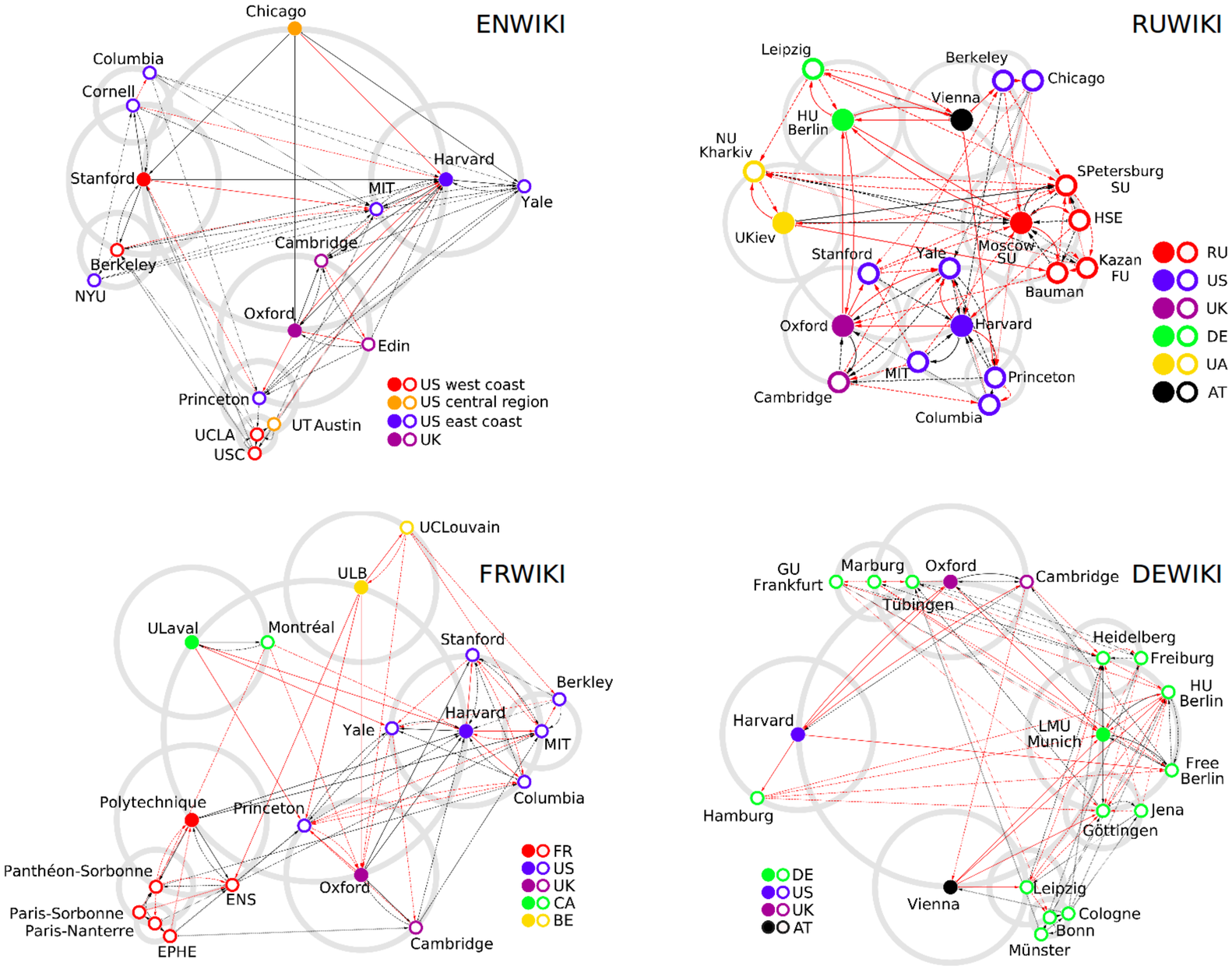




ENWIKI


Chicago
Columbia
Cornell
Stanford
Harvard
MIT
Yale
Cambridge
Berkeley
NYU
Oxford
Edin
Princeton
UT Austin
UCLA
USC
US west coast
US central region
US east coast
UK

Reduced networks of top20 universities



Outline

Influence and interactions of ...

World Universities

Infectious diseases

Cancer and related drugs

... through the prism of Wikipedia

TABLE List of infectious diseases ordered by type (82 of bacterial type, 60 of viral type, 52 of parasitic type, 24 of fungal type, 6 of multiple types, and 2 of other types) then by PageRank of the corresponding article in English edition of Wikipedia . The $n_d = 230$ elements of this list have been extracted from the list of infectious diseases article in 2017 English Wikipedia .

Infectious diseases							
Bacterial type		Bacterial type (Cont.)		Viral type (Cont.)		Parasitic type (Cont.)	
1	Tuberculosis	61	Prevotella	38	CCHF[e]	37	Anaplasmosis
2	Cholera	62	Ehrlichiosis	39	Caliciviridae	38	Baylisascaris
3	Typhoid fever	63	Vibrio vulnif cus	40	PML[f]	39	Fasciolopsiasis
4	Syphilis	64	Actinomycosis	41	Rift Valley fever	40	Head lice infestation
5	Bubonic plague	65	Ureaplasma urealyticum	42	Lassa fever	41	Angiostrongyliasis
6	Leprosy	66	Melioidosis	43	Astrovirus	42	Isosporiasis
7	Sepsis	67	Granuloma inguinale	44	Fifth disease	43	Diphyllobothriasis
8	Meningitis	68	Buruli ulcer	45	VEE[g]	44	Cyclosporiasis
9	Plague (disease)	69	Yersinia pseudotuberculosis	46	Molluscum contagiosum	45	Balantidiasis
10	Typhus	70	Kingella kingae	47	MERS[h]	46	Blastocystosis
11	Diphtheria	71	H. granulocytic anaplasmosis	48	Monkeypox	47	Dientamoebiasis
12	Anthrax	72	Pneumococcal inf.	49	Hand, foot, and mouth disease	48	Gnathostomiasis
13	Helicobacter pylori	73	Nocardiosis	50	Bolivian hemorrhagic fever	49	Capillariasis
14	Lyme disease	74	Bartonellosis	51	H. metapneumovirus	50	Metagonimiasis
15	Staphylococcus	75	Group B streptococcal inf.	52	HFRS[i]	51	Pediculosis corporis
16	Haemophilus inf uenzae	76	Pasteurellosis	53	LCM[j]	52	Free-living Amoebozoa inf.
17	Tetanus	77	Carrion's disease	54	BK virus	fungal type	
18	Gonorrhea	78	Yersiniosis	55	Colorado tick fever	1	Candidiasis
19	Pertussis	79	Mycoplasma pneumonia	56	Argentine hemorrhagic fever	2	Histoplasmosis
20	Botulism	80	HME[a]	57	Heartland virus	3	Athlete's foot
21	Chlamydia inf.	81	Ehrlichiosis ewingii inf.	58	Venezuelan hemorrhagic fever	4	Coccidioidomycosis
22	Rickettsia	82	A. haemolyticum[b]	59	Brazilian hemorrhagic fever	5	Chytridiomycosis
23	Brucellosis	Viral type		60	H. bocavirus	6	Pneumocystis pneumonia
24	Leptospirosis	1	HIV/AIDS	Parasitic type		7	Cryptococcosis
25	Pelvic inf ammatory disease	2	Smallpox	1	Malaria	8	Onychomycosis
26	Cellulitis	3	Inf uenza	2	Schistosomiasis	9	Aspergillosis
27	Legionnaires' disease	4	Measles	3	African trypanosomiasis	10	Tinea versicolor
28	Clostridium diff cile inf.	5	Yellow fever	4	Toxoplasmosis	11	Tinea cruris
29	Rocky Mountain spotted fever	6	Poliomyelitis	5	Onchocerciasis	12	Paracoccidioidomycosis
30	Enterococcus	7	SARS[c]	6	Scabies	13	Tinea corporis
31	Bacterial vaginosis	8	Hepatitis C	7	Chagas disease	14	Blastomycosis
32	Bacterial pneumonia	9	Hepatitis B	8	Filariasis	15	Zygomycosis
33	Epidemic typhus	10	Ebola virus disease	9	Hookworm inf.	16	Tinea capitis
34	Trachoma	11	Common cold	10	Leishmaniasis	17	Tinea manuum
35	Salmonellosis	12	Rabies	11	Trichinosis	18	Sporotrichosis
36	Tularemia	13	Dengue fever	12	Giardiasis	19	Tinea barbae
37	Kawasaki disease	14	Chickenpox	13	Amoebiasis	20	Tinea nigra
38	Bacteroides	15	H. papillomavirus inf.	14	Trichomoniasis	21	Chromoblastomycosis
39	Acinetobacter	16	West Nile fever	15	Cryptosporidiosis	22	White piedra
40	Chlamydophila pneumoniae	17	Herpes simplex	16	Myiasis	23	Black piedra
41	Q fever	18	Rubella	17	Dracunculiasis	24	Geotrichosis
42	Shigellosis	19	Hepatitis A	18	Lymphatic f lariasis	Multiple types	
43	Gas gangrene	20	Cytomegalovirus	19	Echinococcosis	1	Pneumonia
44	Bacillus cereus	21	Mumps	20	Cysticercosis	2	Foodborne illness
45	Group A streptococcal inf.	22	Infectious mononucleosis	21	Babesiosis	3	Hemolytic-uremic syndrome
46	Meningococcal disease	23	Shingles	22	Pinworm inf.	4	Keratitis
47	Burkholderia	24	Norovirus	23	Ascariasis	5	Neonatal conjunctivitis
48	Relapsing fever	25	Rotavirus	24	Strongyloidiasis	6	Mycetoma
49	Glanders	26	Hantavirus	25	Toxocariasis	Prionic type	
50	Psittacosis	27	H. respiratory syncytial virus	26	Anisakis	1	Creutzfeldt–Jakob disease
51	Listeriosis	28	Marburg virus	27	Fasciolosis	2	Kuru (disease)
52	Rickettsialpox	29	Rhinovirus	28	Paragonimiasis	3	Fatal familial insomnia
53	Campylobacteriosis	30	Enterovirus	29	Cutaneous larva migrans	4	GSS[k]
54	Naegleriasis	31	Chikungunya	30	Trichuriasis	Other types	
55	Murine typhus	32	Hepatitis D	31	Hymenolepiasis	1	Desmodesmus
56	Fusobacterium	33	Viral pneumonia	32	Clonorchiasis	2	Rhinosporidiosis
57	Chancroid	34	Hepatitis E	33	Pediculosis pubis		
58	Cat-scratch disease	35	SSPE[d]	34	Microsporidiosis		
59	Staphylococcal inf.	36	Roseola	35	Opisthorchiasis		
60	Vibrio parahaemolyticus	37	H. parainf uenza viruses	36	Taeniasis		

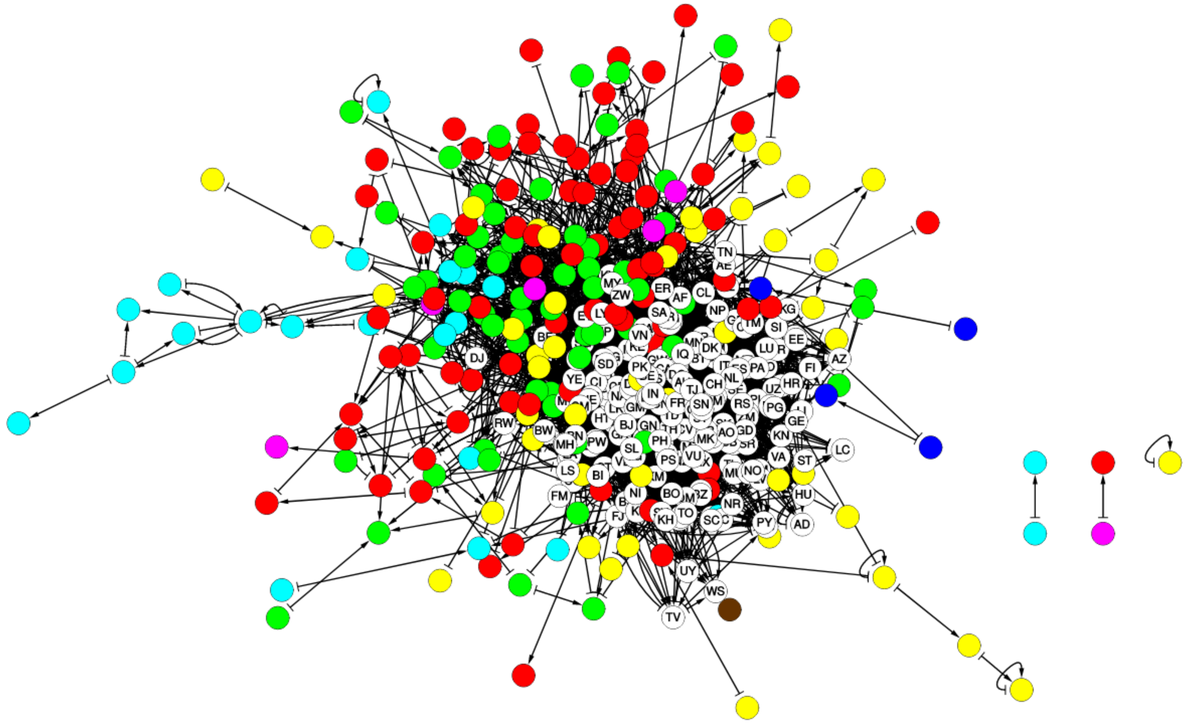
Abbreviations H. and inf. stand for Human and infection. [a] HME: Human monocytotropic ehrlichiosis. [b] A. haemolyticum: Arcanobacterium haemolyticum. [c] SARS: Severe acute respiratory syndrome. [d] SSPE: Subacute sclerosing panencephalitis. [e] CCHF: Crimean–Congo hemorrhagic fever. [f] PML: Progressive multifocal leukoencephalopathy. [g] VEE: Venezuelan equine encephalitis virus. [h] MERS: Middle East respiratory syndrome. [i] HFRS: Hantavirus hemorrhagic fever with renal syndrome. [j] LCM: Lymphocytic choriomeningitis. [k] GSS: Gerstmann–Sträussler–Scheinker syndrome.

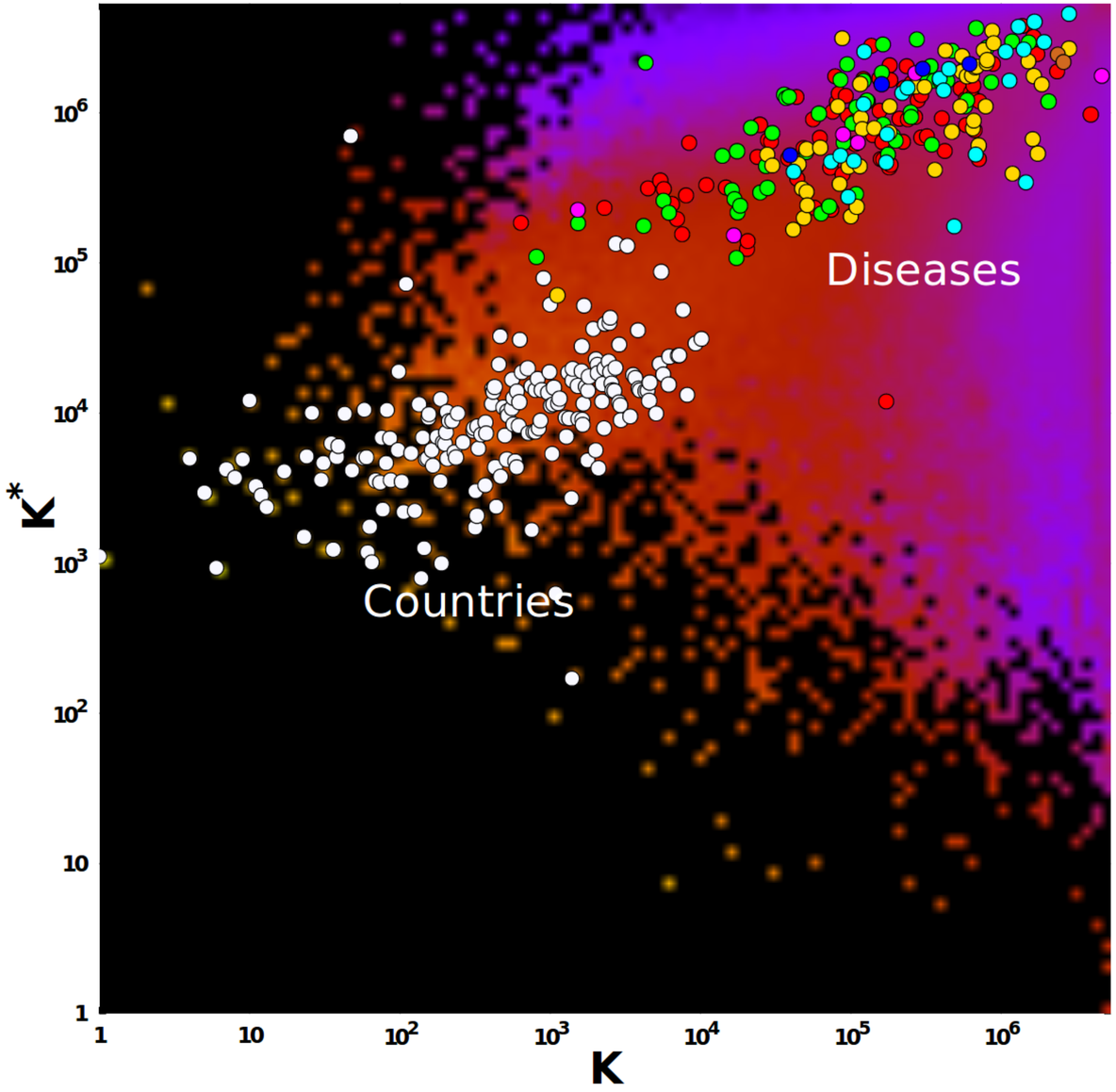
TABLE Ranking of infectious diseases and countries in English Wikipedia 2017 according to PageRank algorithm. The color code distinguishes type of infectious diseases: **bacterial**, **viral**, **parasitic**, **fungal**, **prionic**, **multiple origin**, and **other origin**.

Rank	Disease or country	Rank	Disease or country	Rank	Disease or country	Rank	Disease or country
1	United States	228	Haemophilus inf uenzae	294	Cysticercosis	360	Bolivian hemorrhagic fever
	...	229	Tetanus	295	Babesiosis	361	Blastomycosis
105	Sudan	230	H. papillomavirus inf.	296	Bacteroides	362	Cutaneous larva migrans
106	Tuberculosis	231	West Nile fever	297	Pneumocystis pneumonia	363	H. metapneumovirus
107	Uganda	232	Schistosomiasis	298	Viral pneumonia	364	Zygomycosis
	...	233	Herpes simplex	299	Cryptococcosis	365	Trichuriasis
114	Somalia	234	Gonorrhea	300	Hepatitis E	366	Granuloma inguinale
115	HIV/AIDS	235	Pertussis	301	Acinetobacter	367	Hymenolepiasis
116	Ivory Coast	236	African trypanosomiasis	302	Chlamydophila pneumoniae	368	Clonorchiasis
	...	237	Rubella	303	Q fever	369	HFRS[g]
128	Fiji	238	Hepatitis A	304	Pinworm inf.	370	Buruli ulcer
129	Malaria	239	Cytomegalovirus	305	Shigellosis	371	LCM[h]
130	Mali	240	Botulism	306	Gas gangrene	372	Yersinia pseudotuberculosis
	...	241	Mumps	307	Bacillus cereus	373	Pediculosis pubis
140	Oman	242	Creutzfeldt–Jakob disease	308	Kuru (disease)	374	BK virus
141	Pneumonia	243	Toxoplasmosis	309	SSPE[b]	375	GSS[i]
142	Smallpox	244	Candidiasis	310	Group A streptococcal inf.	376	Kingella kingae
143	Suriname	245	Chlamydia inf.	311	Roseola	377	H. granulocytic anaplasmosis
	...	246	Rickettsia	312	Meningococcal disease	378	Microsporidiosis
162	Malawi	247	Infectious mononucleosis	313	H. parainf uenza viruses	379	Pneumococcal inf.
163	Cholera	248	Onchocerciasis	314	Burkholderia	380	Opisthorchiasis
164	Togo	249	Scabies	315	Onychomycosis	381	Nocardiosis
	...	250	Brucellosis	316	Aspergillosis	382	Taeniasis
185	San Marino	251	Chagas disease	317	CCHF[c]	383	Bartonellosis
186	Inf uenza	252	Shingles	318	Relapsing fever	384	Anaplasmosis
187	Saint Lucia	253	Filariasis	319	Ascariasis	385	Tinea capitis
188	Measles	254	Hookworm inf.	320	Strongyloidiasis	386	Colorado tick fever
189	Palau	255	Leishmaniasis	321	Glanders	387	Baylisascaris
190	Typhoid fever	256	Leptospirosis	322	Psittacosis	388	Fasciolopsiasis
191	Marshall Islands	257	Pelvic inf ammatory disease	323	Listeriosis	389	Group B streptococcal inf.
192	Equatorial Guinea	258	Norovirus	324	Caliciviridae	390	Pasteurellosis
193	Dominica	259	Cellulitis	325	PML[d]	391	Head lice infestation
194	Guinea-Bissau	260	Trichinosis	326	Rickettsialpox	392	Angiostrongyliasis
195	Syphilis	261	Rotavirus	327	Tinea versicolor	393	Isosporiasis
196	Comoros	262	Hantavirus	328	Campylobacteriosis	394	Argentine hemorrhagic fever
197	Djibouti	263	Legionnaires' disease	329	Naegleriasis	395	Diphyllobothriasis
198	Yellow fever	264	Histoplasmosis	330	Murine typhus	396	Heartland virus
199	Bubonic plague	265	Clostridium diff cile inf.	331	Tinea cruris	397	Cyclosporiasis
200	Fed. States of Micronesia	266	Rocky Mountain spotted fever	332	Fusobacterium	398	Carrion's disease
201	Poliomyelitis	267	Enterococcus	333	Rift Valley fever	399	Balantidiasis
202	Tuvalu	268	Bacterial vaginosis	334	Lassa fever	400	Tinea manuum
203	Leprosy	269	Giardiasis	335	Chancroid	401	Sporotrichosis
204	Sepsis	270	Bacterial pneumonia	336	Cat-scratch disease	402	Venezuelan hemorrhagic fever
205	Nauru	271	Amoebiasis	337	Neonatal conjunctivitis	403	Blastocystosis
206	St. Vincent & Grenadines	272	H. respiratory syncytial virus	338	Toxocariasis	404	Tinea barbae
207	Meningitis	273	Athlete's foot	339	Astrovirus	405	Yersiniosis
208	Kiribati	274	Trichomoniasis	340	Fifth disease	406	Tinea nigra
209	Plague (disease)	275	Epidemic typhus	341	Staphylococcal inf.	407	Chromoblastomycosis
210	Saint Kitts and Nevis	276	Hemolytic-uremic syndrome	342	Vibrio parahaemolyticus	408	Dientamoebiasis
211	Typhus	277	Marburg virus	343	Prevotella	409	Brazilian hemorrhagic fever
212	Antigua and Barbuda	278	Trachoma	344	Fatal familial insomnia	410	Gnathostomiasis
213	São Tomé & Príncipe	279	Rhinovirus	345	Anisakis	411	Mycoplasma pneumonia
214	Diphtheria	280	Salmonellosis	346	Ehrlichiosis	412	Capillariasis
215	SARS[a]	281	Coccidioidomycosis	347	VEE[e]	413	White piedra
216	Anthrax	282	Cryptosporidiosis	348	Molluscum contagiosum	414	HME[j]
217	Hepatitis C	283	Myiasis	349	MERS[f]	415	Metagonimiasis
218	Foodborne illness	284	Enterovirus	350	Monkeypox	416	Pediculosis corporis
219	Hepatitis B	285	Chytridiomycosis	351	Fasciolosis	417	Black piedra
220	Ebola virus disease	286	Tularemia	352	Paracoccidioidomycosis	418	H. bocavirus
221	Common cold	287	Kawasaki disease	353	Hand, foot, and mouth disease	419	Ehrlichiosis ewingii inf.
222	Rabies	288	Chikungunya	354	Vibrio vulnif cus	420	Desmodesmus
223	Dengue fever	289	Hepatitis D	355	Actinomycosis	421	Rhinosporidiosis
224	Helicobacter pylori	290	Dracunculiasis	356	Ureaplasma urealyticum	422	Free-living Amoebozoa inf.
225	Lyme disease	291	Keratitis	357	Tinea corporis	423	Geotrichosis
226	Chickenpox	292	Lymphatic f lariasis	358	Melioidosis	424	A. haemolyticum[k]
227	Staphylococcus	293	Echinococcosis	359	Paragonimiasis	425	Mycetoma

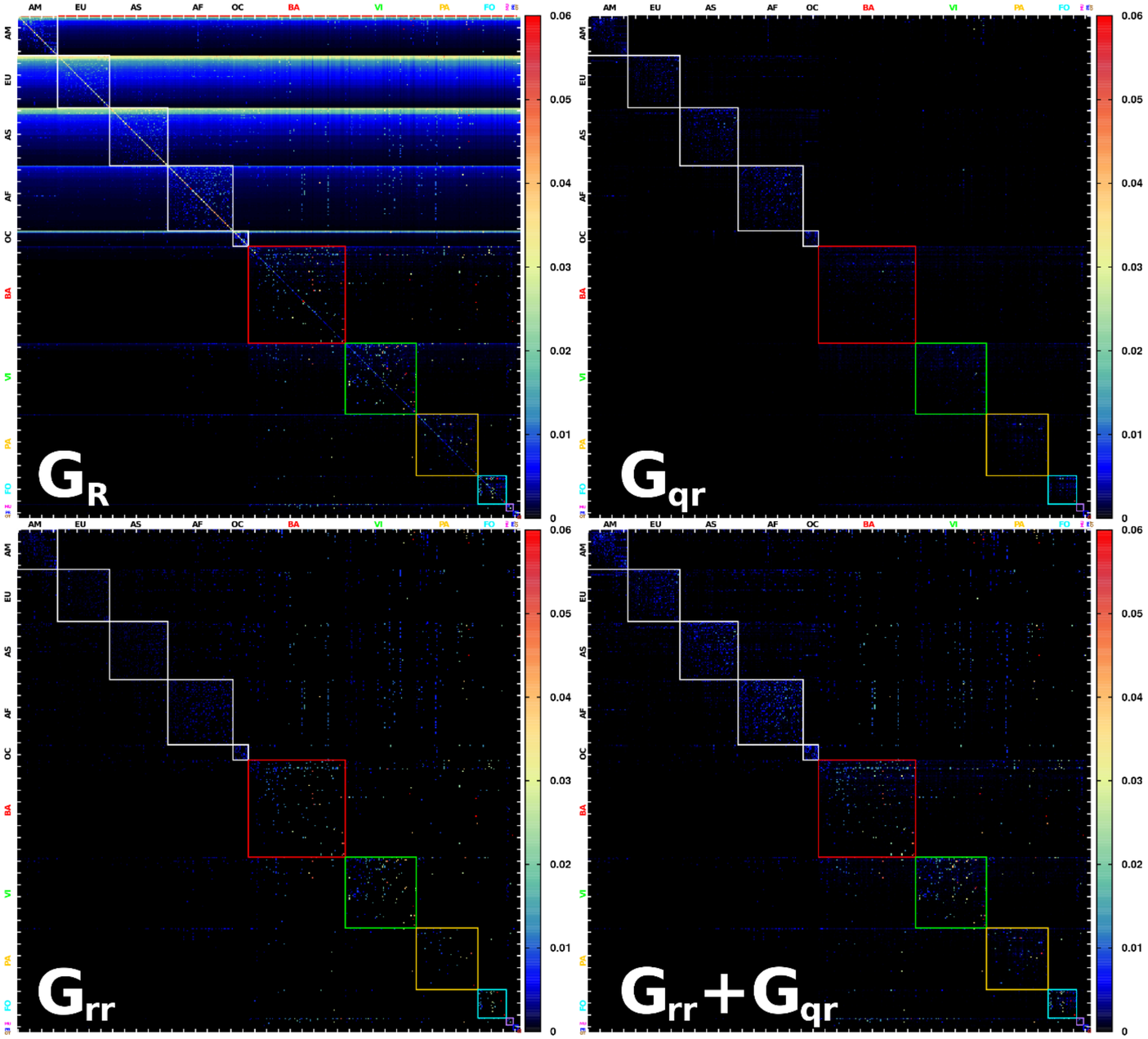
Abbreviations H. and inf. stand for Human and infection. [a] SARS: Severe acute respiratory syndrome. [b] SSPE: Subacute sclerosing panencephalitis. [c] CCHF: Crimean–Congo hemorrhagic fever. [d] PML: Progressive multifocal leukoencephalopathy. [e] VEE: Venezuelan equine encephalitis virus. [f] MERS: Middle East respiratory syndrome. [g] HFRS: Hantavirus hemorrhagic fever with renal syndrome. [h] LCM: Lymphocytic choriomeningitis. [i] GSS: Gerstmann–Sträussler–Scheinker syndrome. [j] HME: Human monocytotropic ehrlichiosis. [k] A. haemolyticum: Arcanobacterium haemolyticum.

"Infectious diseases and countries" sub-network in English Wikipedia

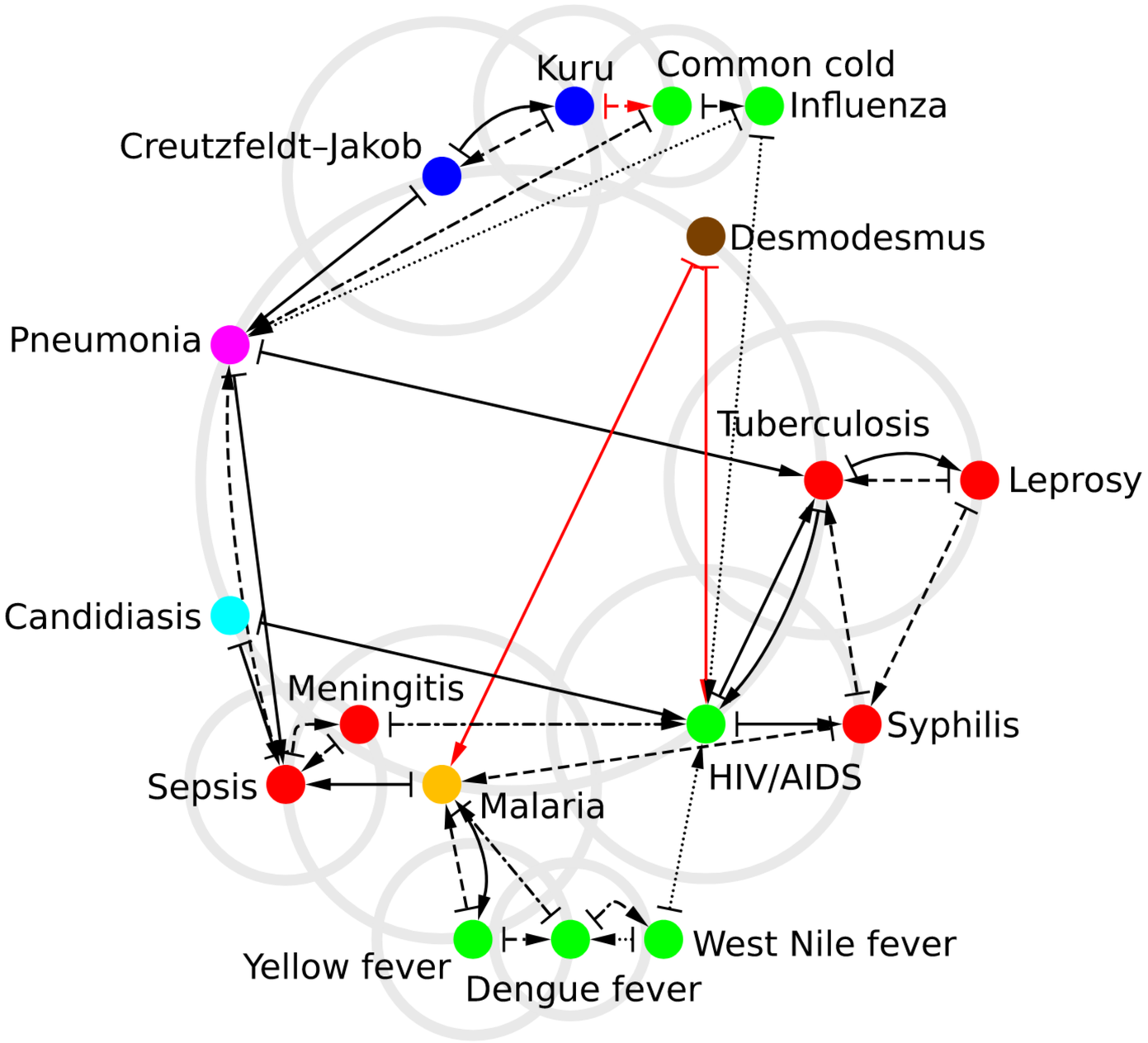


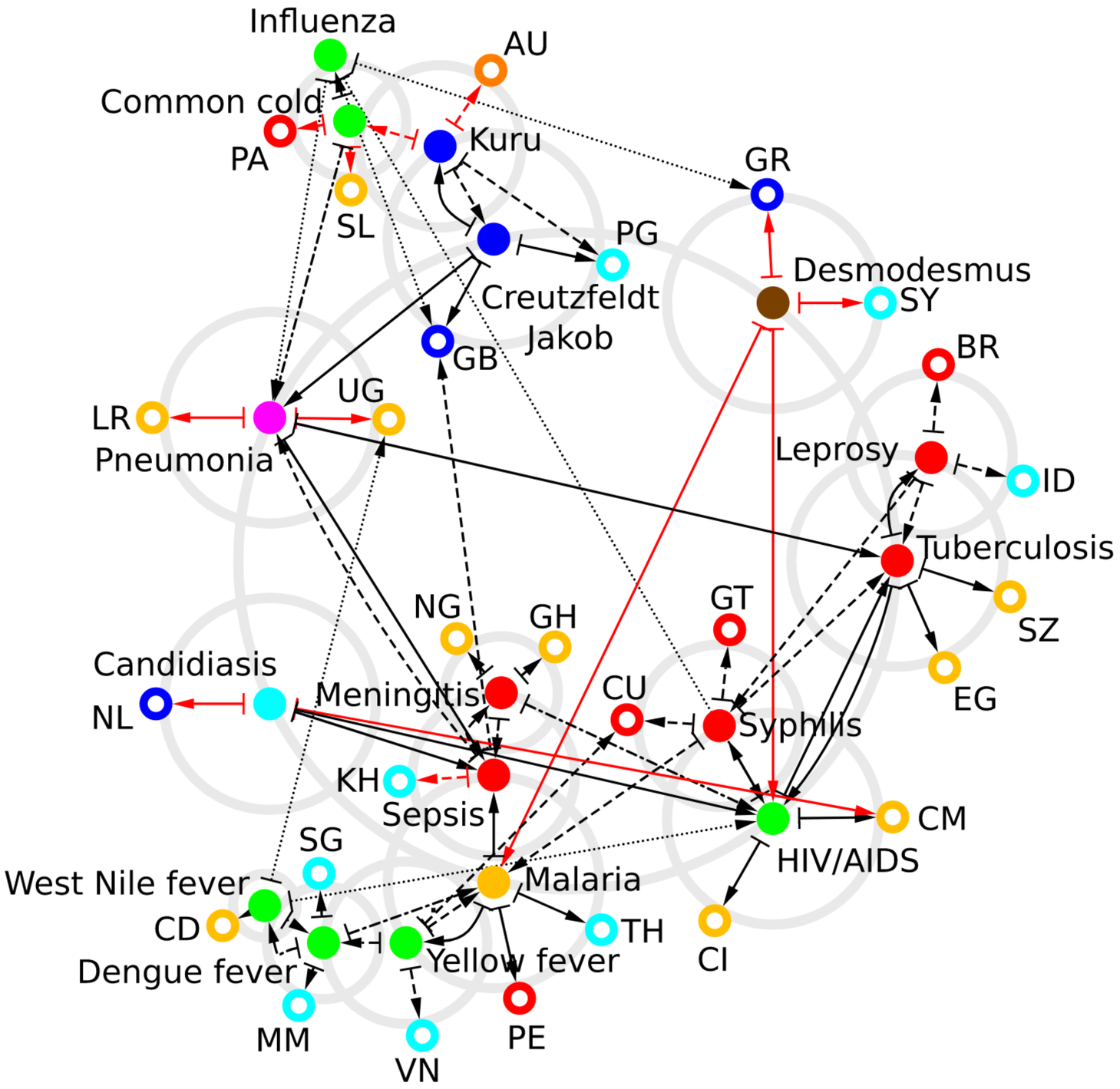


Reduced Google matrix for "infectious diseases and countries"

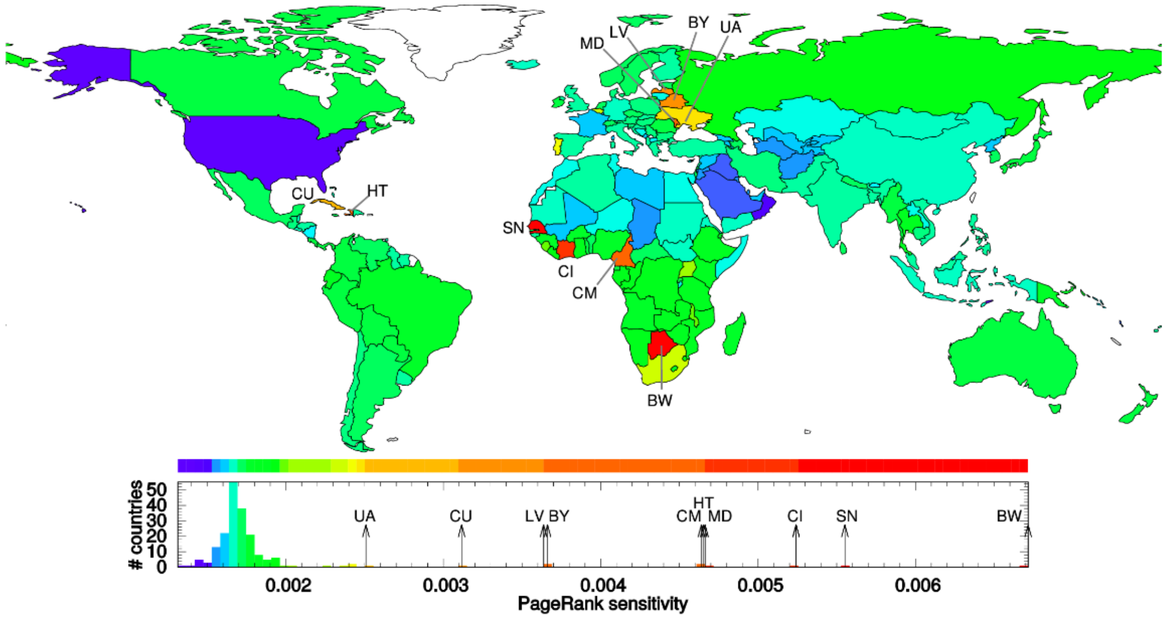


Infectious diseases reduced network

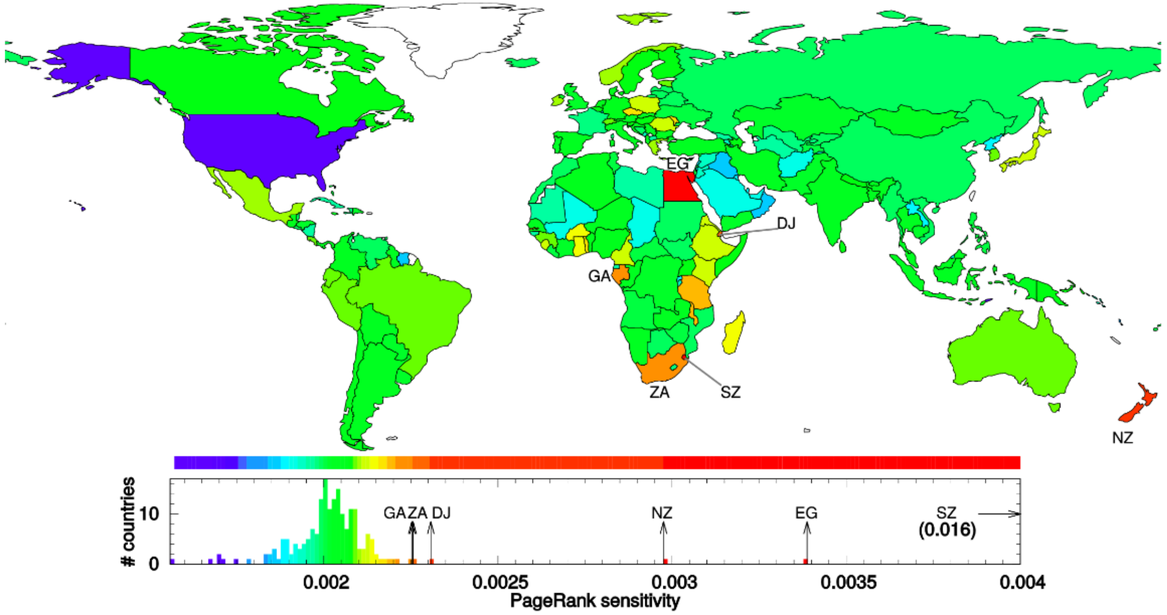



Influenza
AU
Common cold
PA
Kuru
SL
GR
PG
Desmodesmus
SY
Creutzfeldt
Jakob
GB
BR
UG
LR
Pneumonia
Leprosy
ID
Tuberculosis
SZ
NG
GH
GT
EG
Candidiasis
NL
Meningitis
CU
Syphilis
KH
Sepsis
CM
SG
HIV/AIDS
West Nile fever
Malaria
CD
TH
CI
Dengue fever
Yellow fever
MM
VN
PE

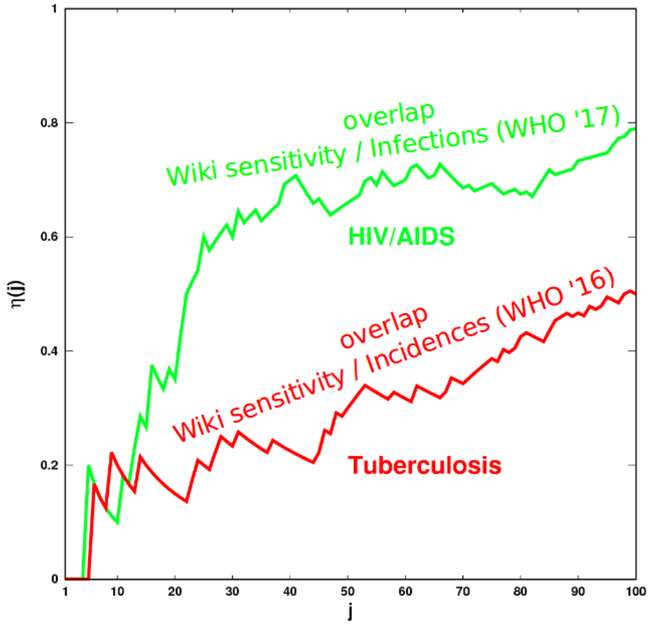
$D(\text{HIV} \rightarrow \text{Country}, \text{Country})$



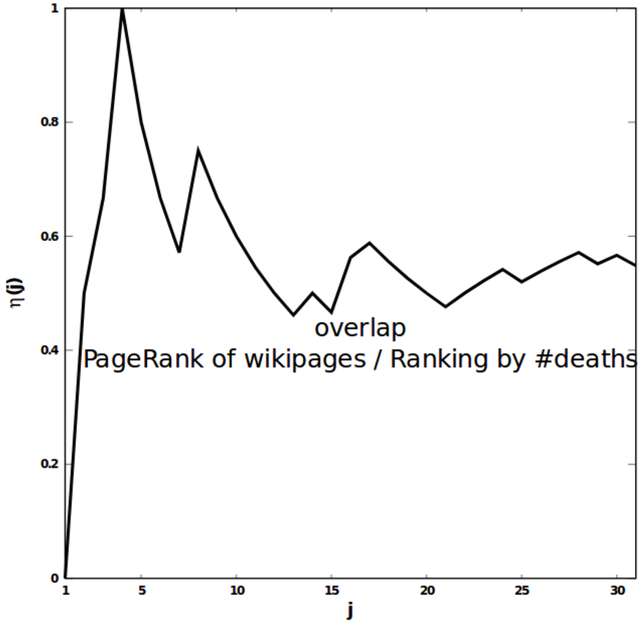
$D(\text{Tuberculosis} \rightarrow \text{Country}, \text{Country})$

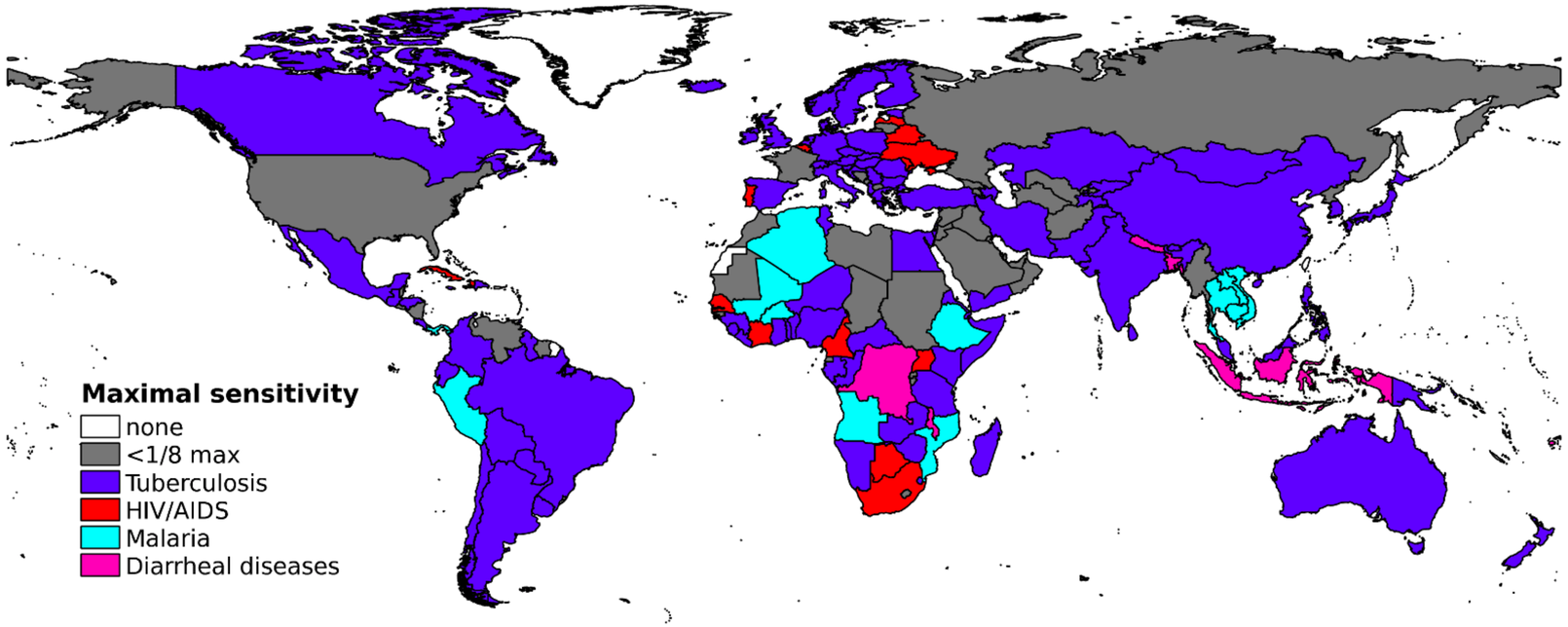


Ranking of countries



Ranking of diseases





Outline

Influence and interactions of ...

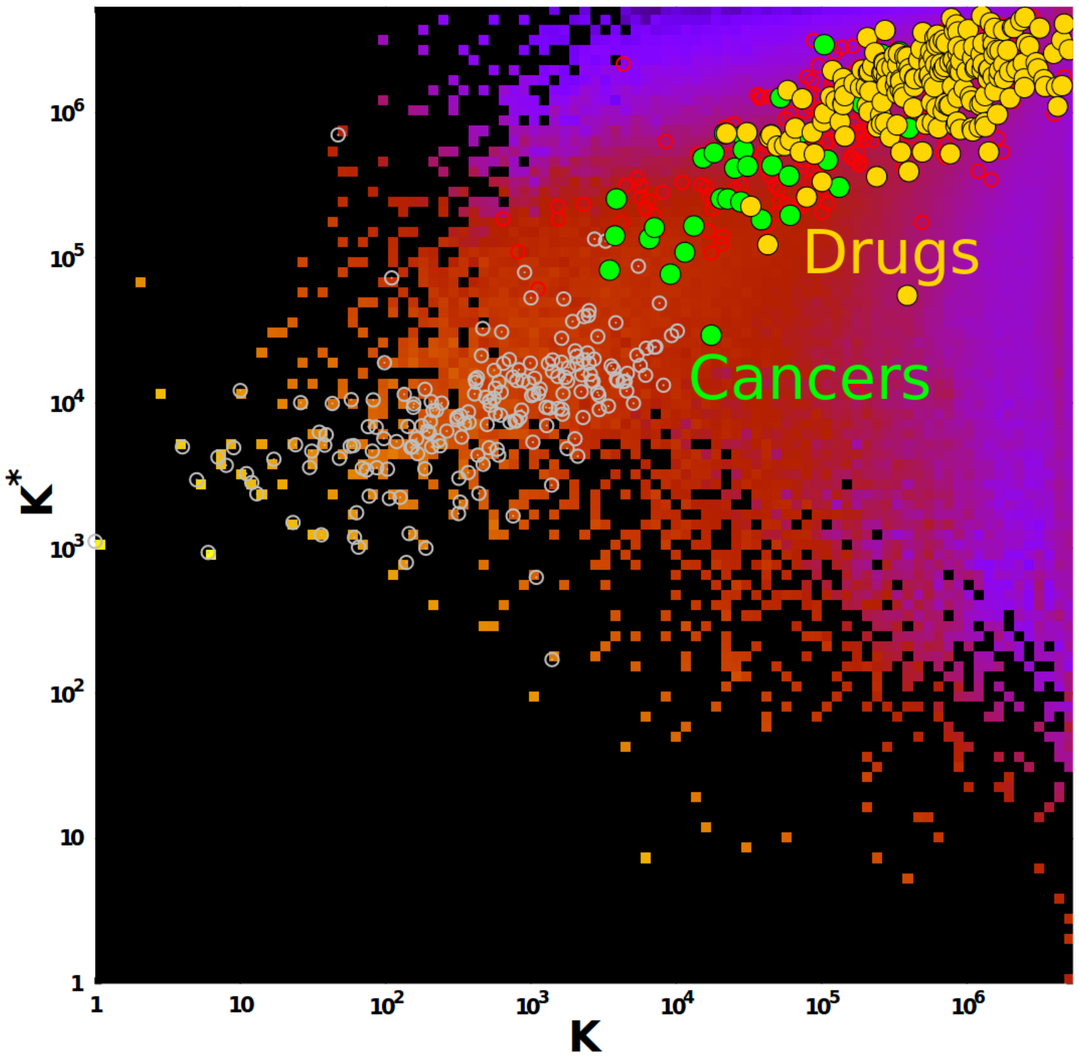
World Universities

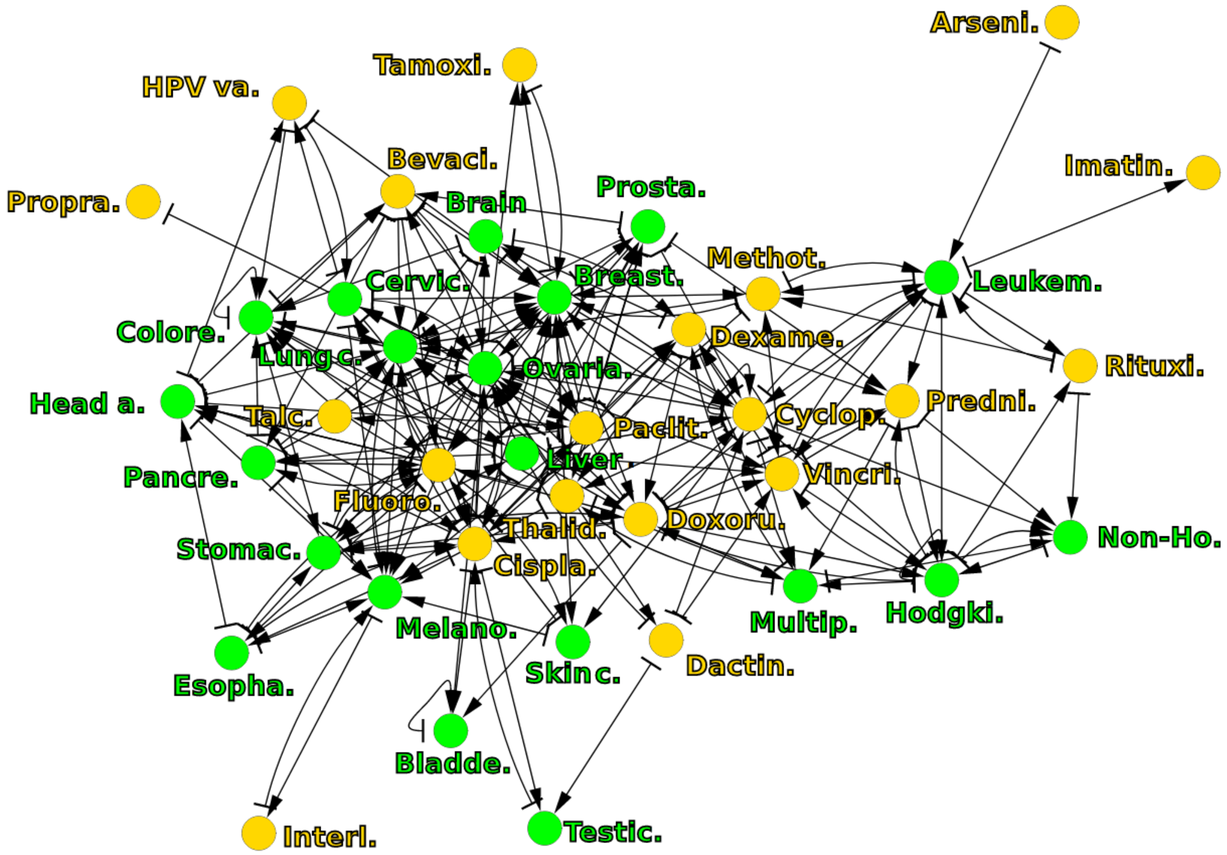
Infectious diseases

Cancer and related drugs (preliminary results)

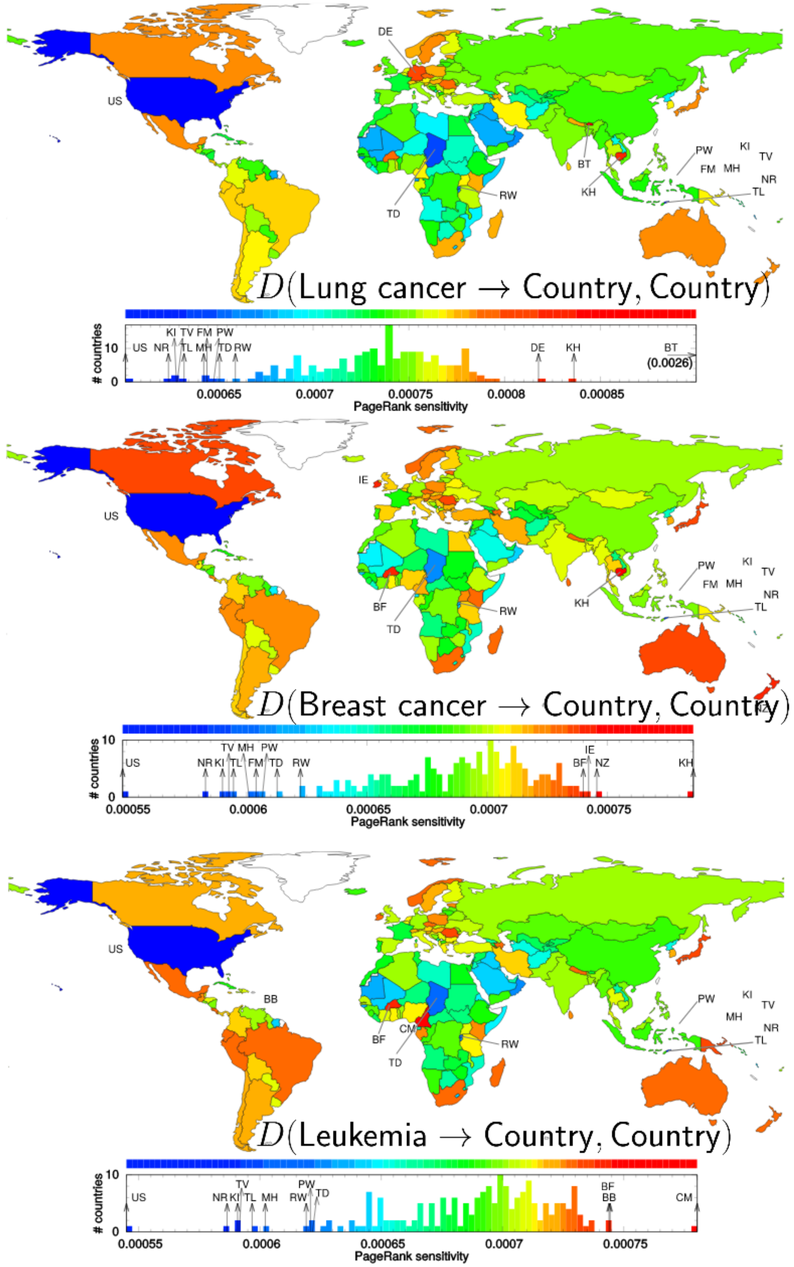
... through the prism of Wikipedia

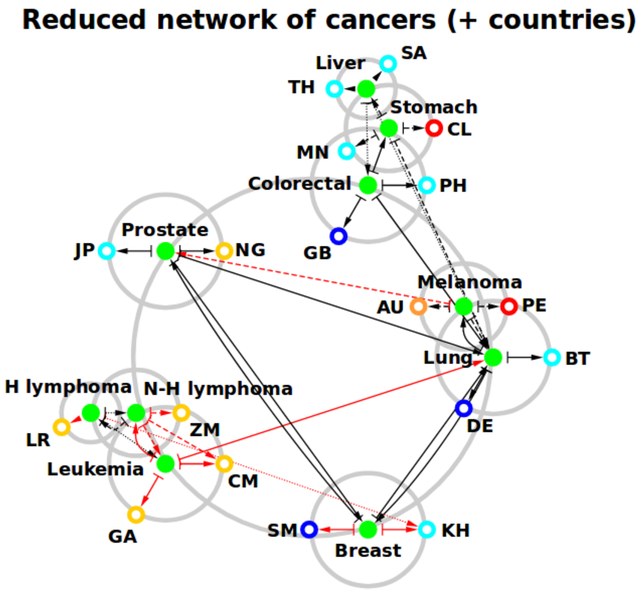
Cancers (37) and related drugs (203) in 2017 English Wikipedia

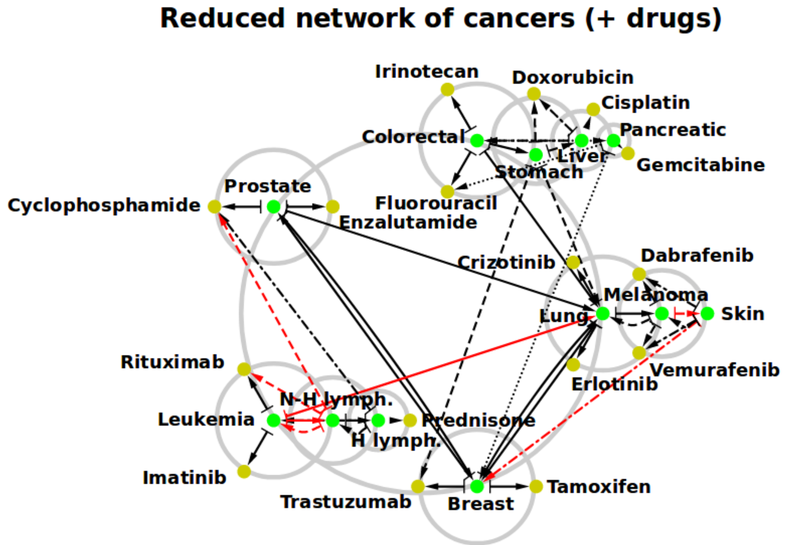




Top 20 cancers and top 20 drugs network in Wikipedia


D(Lung cancer → Country, Country)
countries
PageRank sensitivity
D(Breast cancer → Country, Country)
countries
PageRank sensitivity
D(Leukemia → Country, Country)
countries
PageRank sensitivity


Reduced network of cancers (+ countries)
Liver
SA
TH
Stomach
CL
MN
Colorectal
PH
Prostate
JP
NG
GB
Melanoma
AU
PE
Lung
BT
H lymphoma
N-H lymphoma
LR
ZM
DE
Leukemia
CM
GA
SM
Breast
KH


Reduced network of cancers (+ drugs)
Irinotecan
Doxorubicin
Cisplatin
Colorectal
Pancreatic
Liver
Stomach
Gemcitabine
Prostate
Cyclophosphamide
Fluorouracil
Enzalutamide
Crizotinib
Dabrafenib
Melanoma
Lung
Skin
Rituximab
Vemurafenib
Erlotinib
N-H lymph.
Leukemia
Prednisone
H lymph.
Imatinib
Trastuzumab
Breast
Tamoxifen

Conclusions

Encoding the information of the whole global network, the **Reduced Google matrix** method allows to assign a given weight to each directional relationship between entities of interest.

Moreover **hidden links** can be inferred.

From Wikipedia, one can **unveil effective interactions between entities of interest** (universities, countries , infectious diseases, cancers, drugs, ...)

Inverse problem: for a given University in Wikipedia, the reduced Google matrix method could be used to find articles which could increase the influence of the University in Wikipedia.

For infectious diseases, results are in accordance with known data and World Health Organization data.

Wikipedia is an encylopedia quite easy to surf for Human, what about other non Human readable datasets, e.g. :

- World Trade Network (see Leo Ermann's talk)

- protein-protein causal interactions networks obtained from curation of huge amount of scientifical articles (see Andrei Zinovyev's talk)

Thank You !

Articles and datasets availble at:

http://perso.utinam.cnrs.fr/~lages/datasets/WRWU/
http://perso.utinam.cnrs.fr/~lages/datasets/WRWU17/
http://perso.utinam.cnrs.fr/~lages/datasets/Wiki4InfectiousDiseases/

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