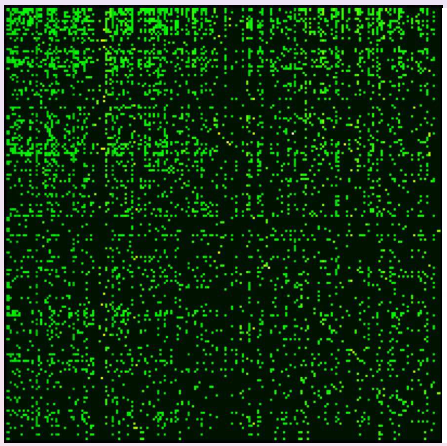
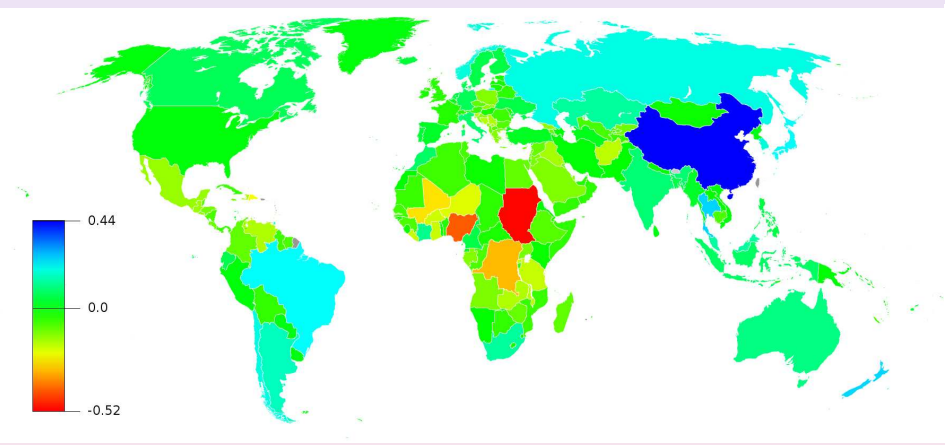
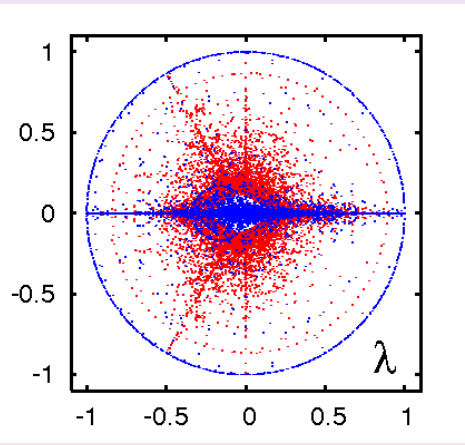
Google matrix: fundamentals and applications

Dima Shepelyansky (Quantware group, LPT CNRS, Toulouse)
www.quantware.ups-tlse.fr/dima

with L.Ermann (CNEA TANDAR), **Klaus Frahm (LPT UPS)**, K.Jaffres-Runser (N7 Tlse), J.Lages (U Besancon); A.Chepelianskii (LPS Orsay)



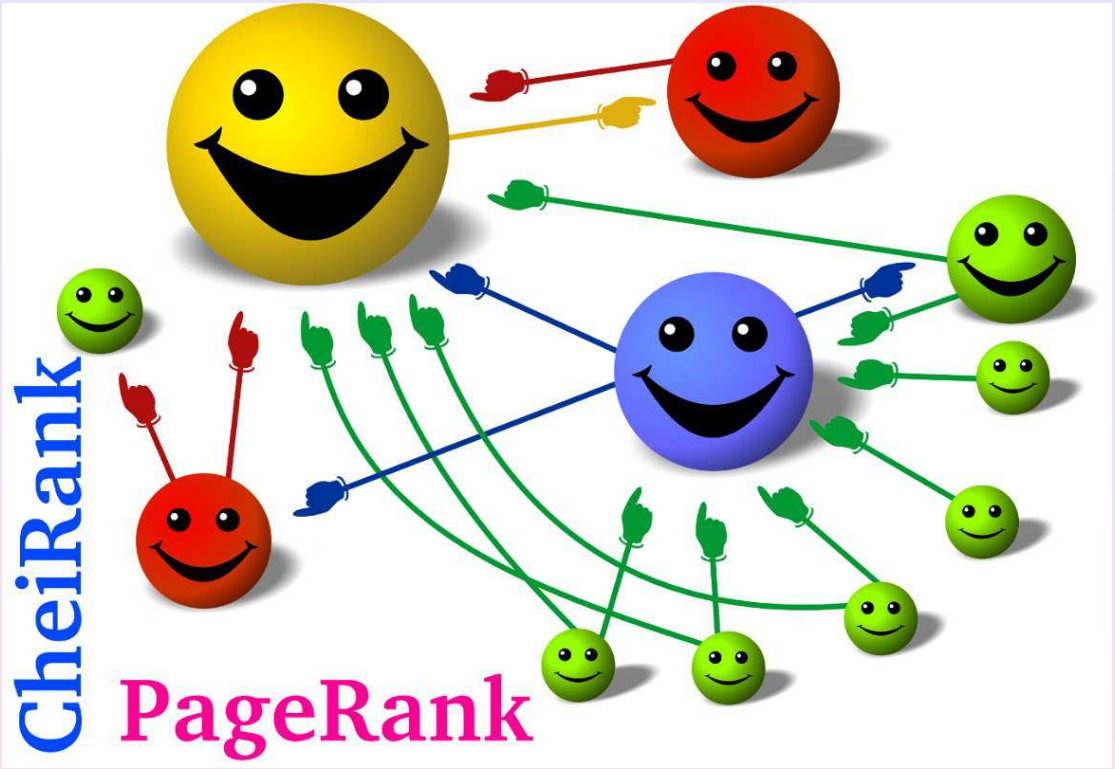


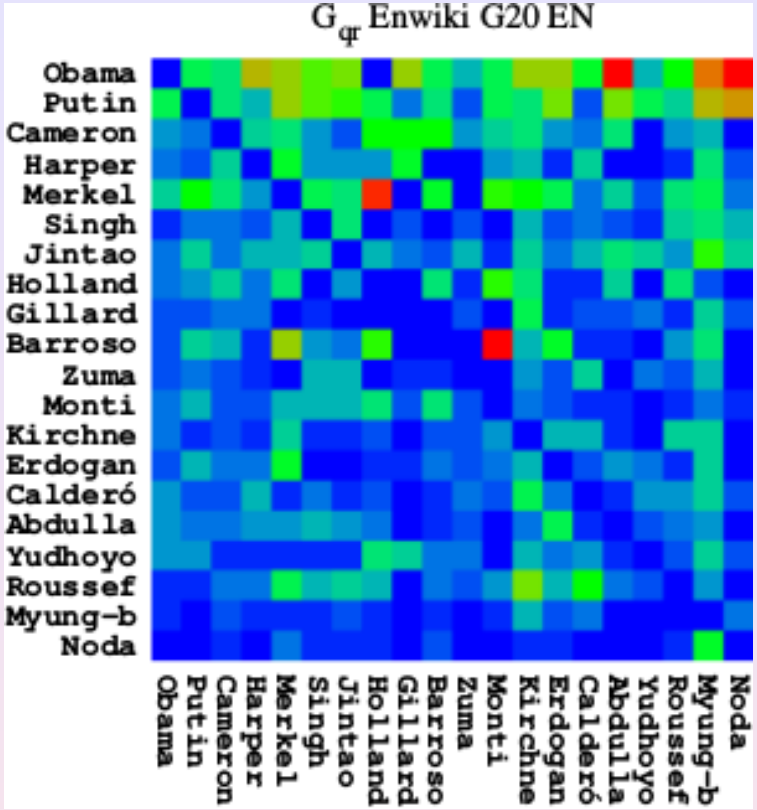


* Markov (1906) $\rightarrow$ Brin and Page (1998) $\rightarrow$ Google matrix and search engines
* reduced Google matrix of directed networks (brief introduction)
* Applications: multiproduct world trade networks (UN COMTRADE), Wikipedia Ranking of World Universitis (WRWU), world terror networks, infectious diseases influence, protein cancer networks ...

Support: EC FET Open NADINE (2012-2015), APLIGOOGLE-CNRS (2016-2017), LABEX NEXT (2017-2019) + thanks to UN COMTRADE

(1906) Markov vs Wigner (1955)





1945: Nuclear physics $\rightarrow$ Wigner (1955)$\rightarrow$ Random Matrix Theory

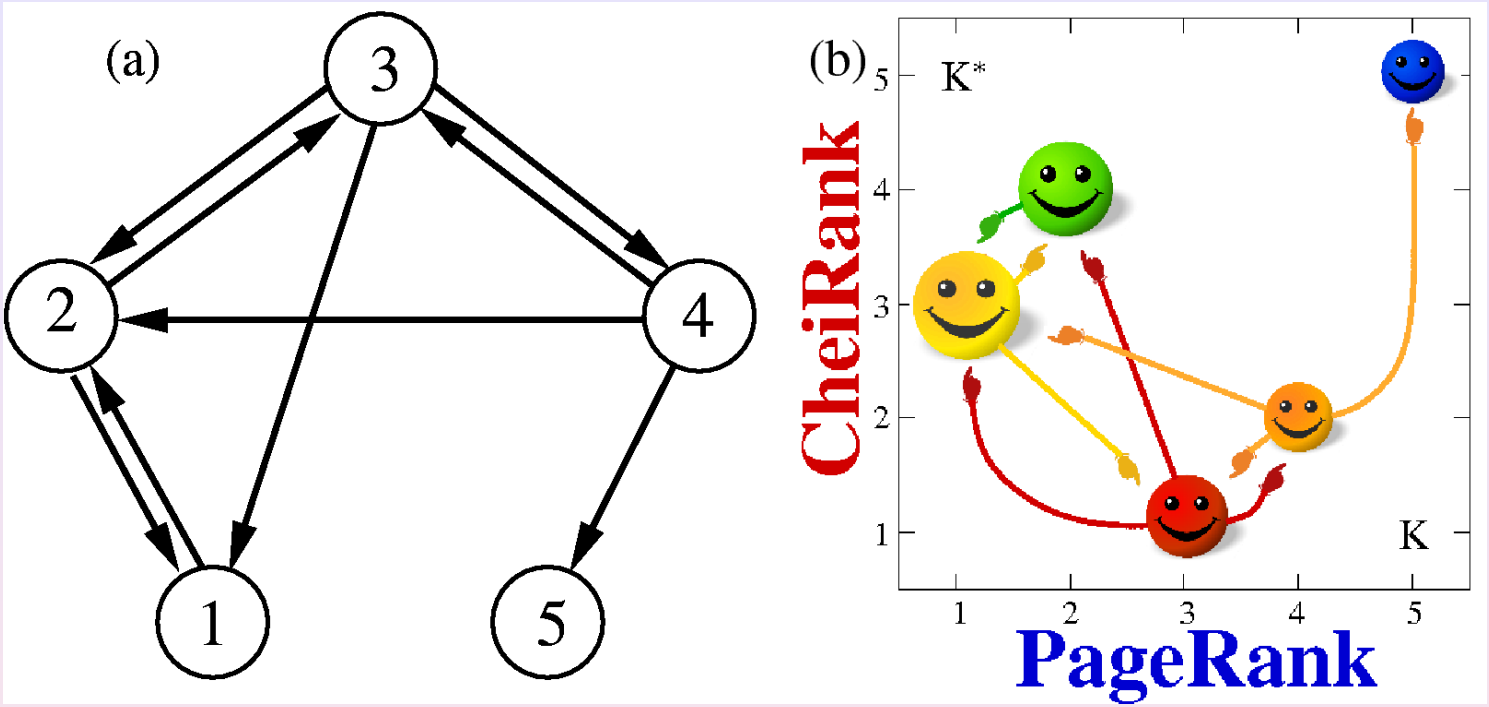
1991: WWW, small world social networks $\rightarrow$ Markov (1906) $\rightarrow$ Google matrix

Despite the importance of large-scale search engines on the web, very little academic research has been done on them.

S.Brin and L.Page, Comp. Networks ISDN Systems **30**, 107 (1998)

Google matrix construction rules

Markov chains (1906) and Directed networks



For a directed network with N nodes the adjacency matrix $\mathbf{A}$ is defined as $A_{ij} = 1$ if there is a link from node j to node i and $A_{ij} = 0$ otherwise. The weighted adjacency matrix is

$$S_{ij} = A_{ij} / \sum_k A_{kj}$$

In addition the elements of columns with only zeros elements are replaced by $1/N$.

Google matrix construction rules

Google Matrix and Computation of PageRank

P $= \mathbf{SP} \Rightarrow$ **P**= stationary vector of **S**; can be computed by iteration of **S**.

To remove convergence problems:

- Replace columns of 0 (dangling nodes) by $\frac{1}{N}$:

$$S = \begin{pmatrix} 0 & 1/2 & 1/3 & 0 & 1/5 \\ 1 & 0 & 1/3 & 1/3 & 1/5 \\ 0 & 1/2 & 0 & 1/3 & 1/5 \\ 0 & 0 & 1/3 & 0 & 1/5 \\ 0 & 0 & 0 & 1/3 & 1/5 \end{pmatrix} \qquad S^* = \begin{pmatrix} 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/2 & 1/3 & 0 & 1 & 0 \\ 0 & 1/3 & 1/2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

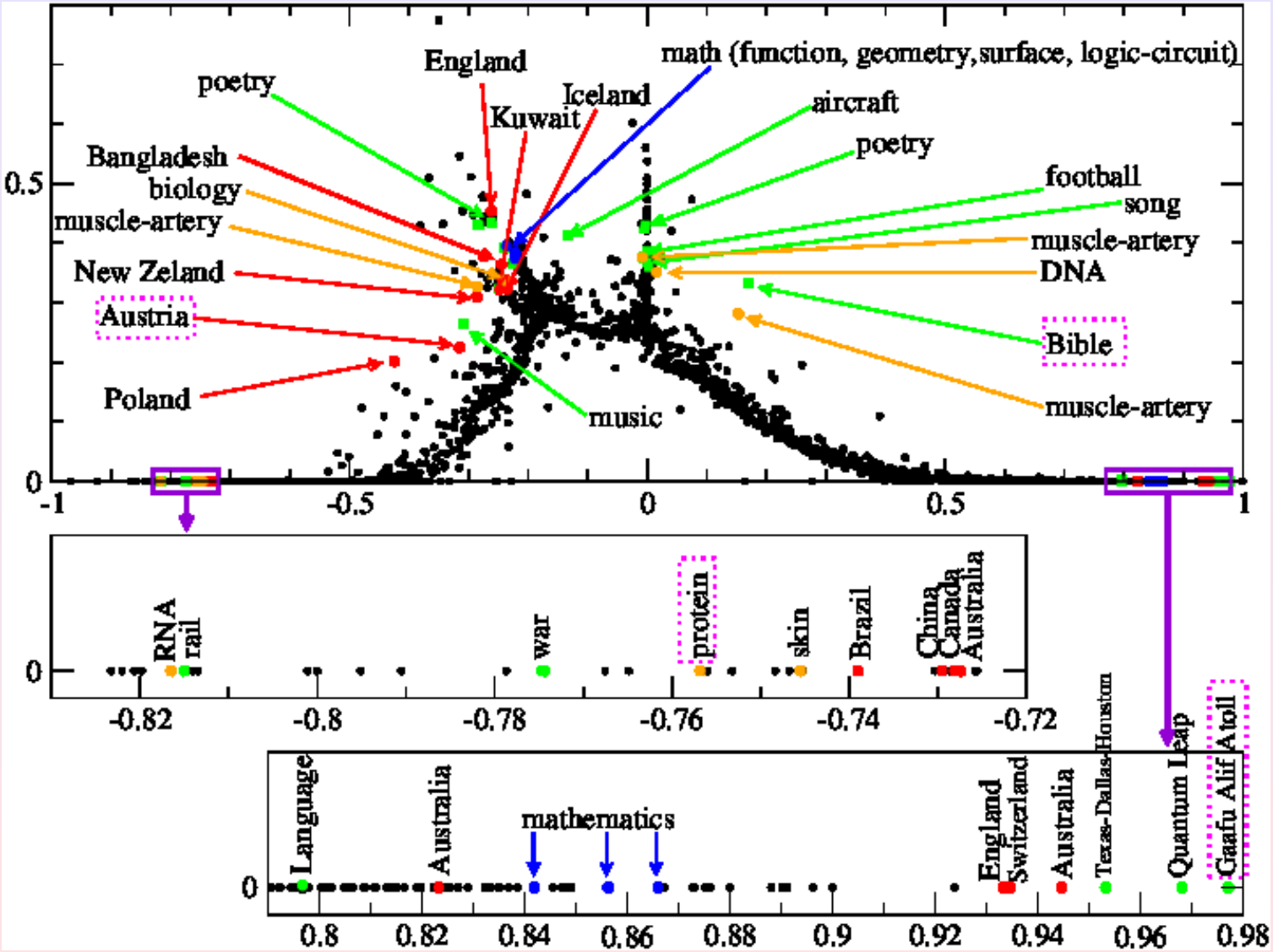
- To remove degeneracies of $\lambda = 1$, replace **S** by **Google matrix**
  $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\frac{\mathbf{E}}{N}$; $GP = \lambda P$ => Perron-Frobenius operator
- α models a random surfer with a random jump after approximately 6 clicks (usually $\alpha = 0.85$); PageRank vector => P at $\lambda = 1$ ($\sum_j P_j = 1$).
- CheiRank vector P^*: $G^* = \alpha\mathbf{S}^* + (1-\alpha)\frac{\mathbf{E}}{N}$, $G^*P^* = P^*$
  (**S*** with inverted link directions)
  Chepelianskii arXiv:1003.5455 (2010) ...

Computation algorithms

* PageRank vector by power iteration:
multiplication of initial random vector by G matrix; convergence to $\lambda = 1$ eigenvector as α^t, about $t = 200$ iterations are enough for double precision convergence (all eigenvalues have $|\lambda| \leq \alpha < 1$ except $\lambda = 1$);
on average there are only about 10-20 nonzero links for each node (about 20 multiplications of vector by a line of matrix)
$\rightarrow$ small-world structure of real networks or six degrees of separation (Milgram Psychology Today (1967));
* Arnoldi algorithm: eigenvalues with largest $|\lambda|$ and related selected eigenvectors corresponding to quasi-isolated communities.
* Reduced Google matrix: description of interactions of subset of selected nodes in a huge network
* GPU codes for reduced Google matrix: 100 times acceleration compared to one-processor computer; collaboration with
Denis Demidov (Russian Academy of Sciences, Kazan; see https://github.com/ddemidov for GPU oriented codes)
* IT companies in Russia:
http://www.gradient-geo.com/en/index.php (Kazan, geo-oil);
http://ledas.com/ (Novosibirsk)

Computation algorithms: spectrum of G

example: spectrum of G for ENWikipedia 2009 (3.28M nodes)



Ermann, Frahm, DS EPJB (2013)

Directed networks analyzed

* Wikipedia editions: EN (2009) $N = 3282257$;
24 editions Wiki2013: $N = 4212493$ EN, $N = 1532978$ DE, $N = 1352825$ FR
24 editions Wiki2017: $N = 5416537$ EN, $N = 2057898$ DE, $N = 1866546$ FR

* Entier Twitter (2009): $N = 41$millions

* Entier Phys. Rev. citation network(1893-2009): $N = 460422$

* World Trade Network (WTN) from UN COMTRADE about 50 years: $N = 227$ for all commodities; multiproduct trade with 61 products $N = 13847$; available with 5000 products and $N \approx 1 million$

* Bitcion network transactions (beginning 2009 till April 2013): $N = 6297009$

* Linux Kernel network: $N = 285509$

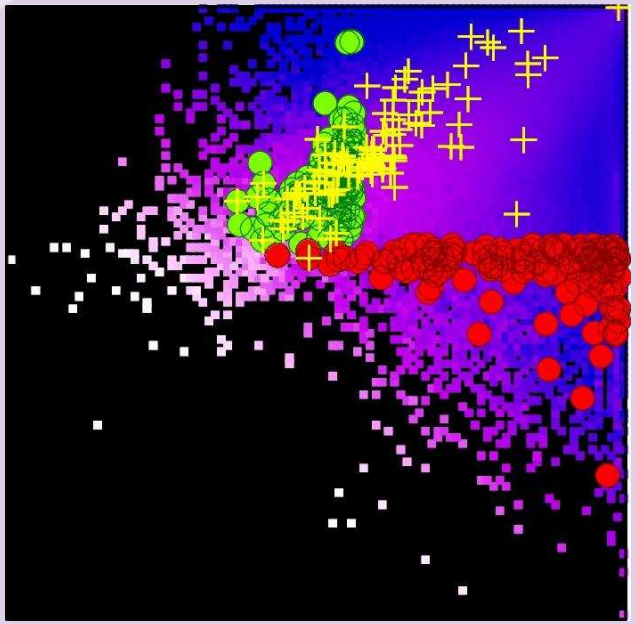
* UK university networks till 2006: U Oxford, Cambridge $N \approx 200000$

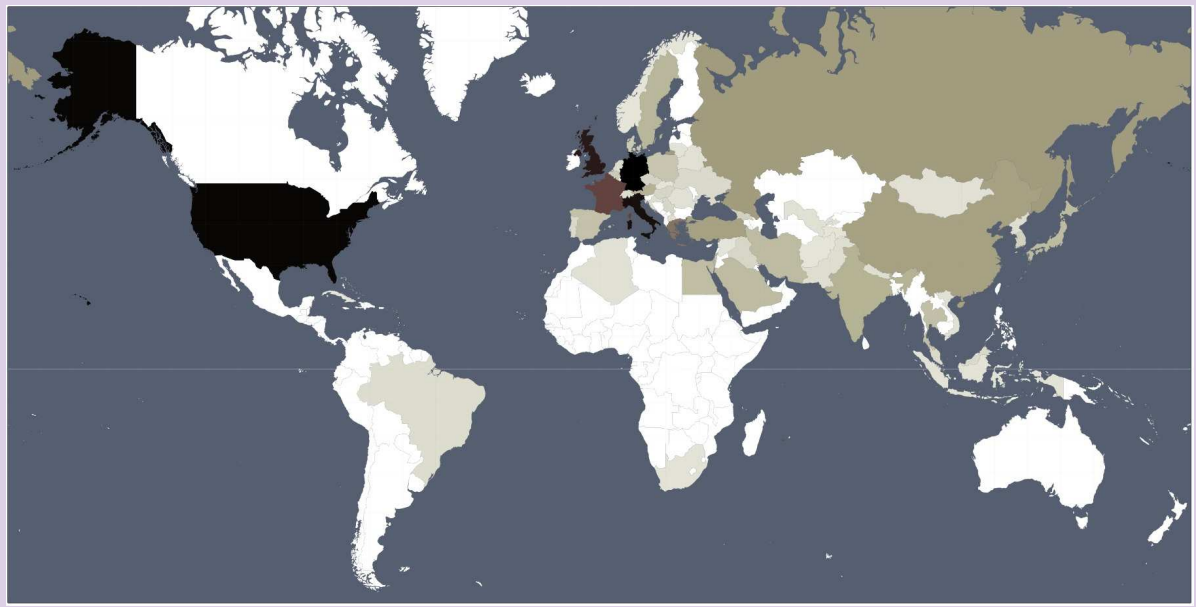
* Network of protein-protein interactions for cancer: $N \approx 3000$

see Ermann, Frahm, DS Rev Mod Phys (2015)
http://www.quantware.ups-tlse.fr/dima/subjgoogle.html

Top historical figures of 24 Wikipedia editions

2DRanking of Wikipedia articles; top 100 historical figures;
comparison with historical studies of M.Hart (37 and 43 percent overlap)
35 centures and all countries by birth place; 17 millions wiki-articles

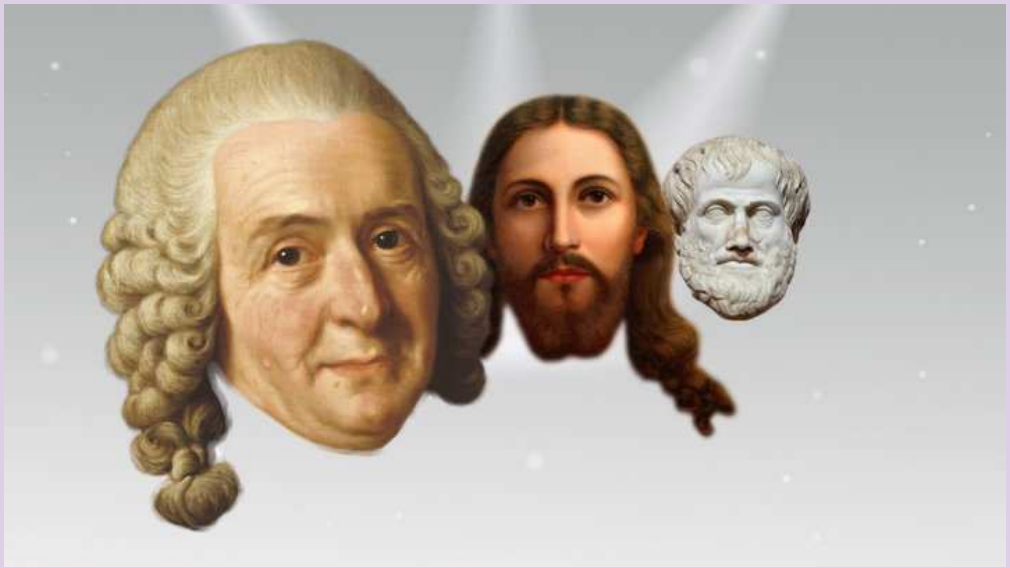




A.Zhirov, O.Zhirov, DS EPJB (2010);
Y.-H.Eom,P.Aragon, D.Laniado, A.Kaltenbrunner, S.Vigna, DS PLoS ONE (2015)

Top historical figures of 24 Wikipedia editions

Top global PageRank historical figures: Carl Linnaeus, Jesus, Aristotle ...



Media highlights: The Guardian, The Independent, The Washington Post, France24, EC CORDIS

==>Uppsala Universitet: “Carl Linnaeus ranked most influential person of all time” ... (about 20 countries)

Competitors: MIT Pantheon project http://pantheon.media.mit.edu (2014); Stony-Brook NY http://www.whoisbigger.com/ (2014)

Reduced Google matrix

A selected network of interest with $N_r < N$ nodes called reduced network.
Block structure of G matrix:

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

with s index for scattering network $N_s = N - N_r$.
Reduced G_R matrix
(analogy with quantum chaotic scattering, nuclear and mesoscopic physics)

$$G_{\rm R} P_r = P_r \quad , \quad G_{\rm R} = G_{rr} + G_{rs}(\mathbf{1} - G_{ss})^{-1} G_{sr} = G_{rr} + G_{pr} + G_{qr}$$

Useful expansion

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

with projector $\mathcal{P}_c = \psi_R \psi_L^T$ on eigenstate of maximal eigenvalue λ_c of G_{ss}, the complementary projector $\mathcal{Q}_c = \mathbf{1} - \mathcal{P}_c$ and $\bar{G}_{ss} = \mathcal{Q}_c G_{ss} \mathcal{Q}_c$.

K.Frahm, DS arxiv:1602.02394 (2016)

Protein-protein interactions for cancer networks

PLOS ONE | Reduced Google matrix for directed biological networks

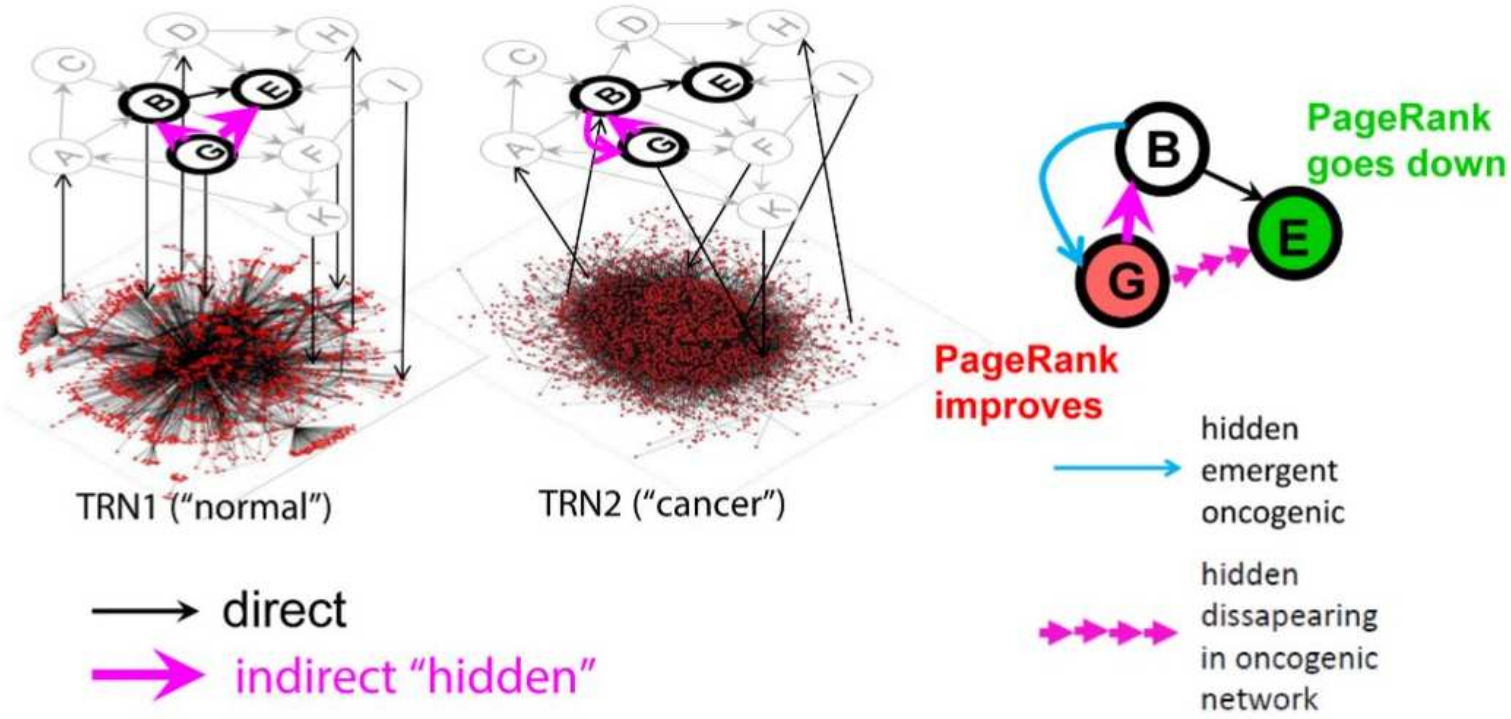


Fig 1. Using reduced Google matrix approach for inferring hidden causal relations in signaling pathways. Here the structure of the context-dependent global regulatory network is symbolically shown as consisting of two layers: the upper (nodes A-K) is the global signaling network whose structure does not depend on the context and the lower is a symbolic view of the contextual transcriptional regulatory network (TRN) whose structure can change between a "normal" and a "cancer" cell. Thick node borders denote a pathway embedded into the global signaling network. Black arrows denote direct physical interactions. Pink arrows denote inferred hidden directed regulations through the global regulatory network (both layers). In the final representation of the pathway (on the right), one can show those hidden regulations which emerge or disappear due to the changes in the TRN structure. Also, the color of the pathway nodes can show the direction of PageRank change: green corresponds to the PageRank decreased in the cancer network while red corresponds to the opposite.

https://doi.org/10.1371/journal.pone.0190812.g001

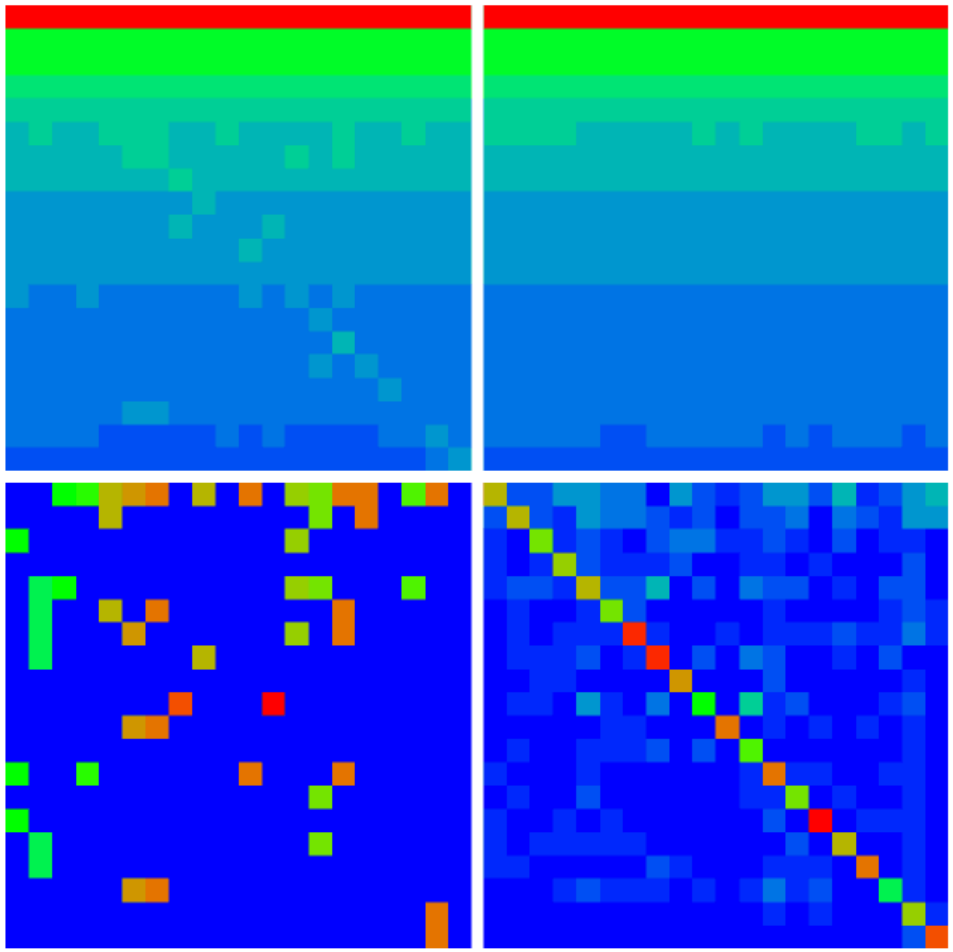
Lages, DS, Zinovyev (PLOS ONE 2018)

G-reduced: G20 political leaders 2012

G_R example: G20 political leaders 2012 from EnWiki2013

$$1 - \lambda_c = 2.465 \times 10^{-4}$$

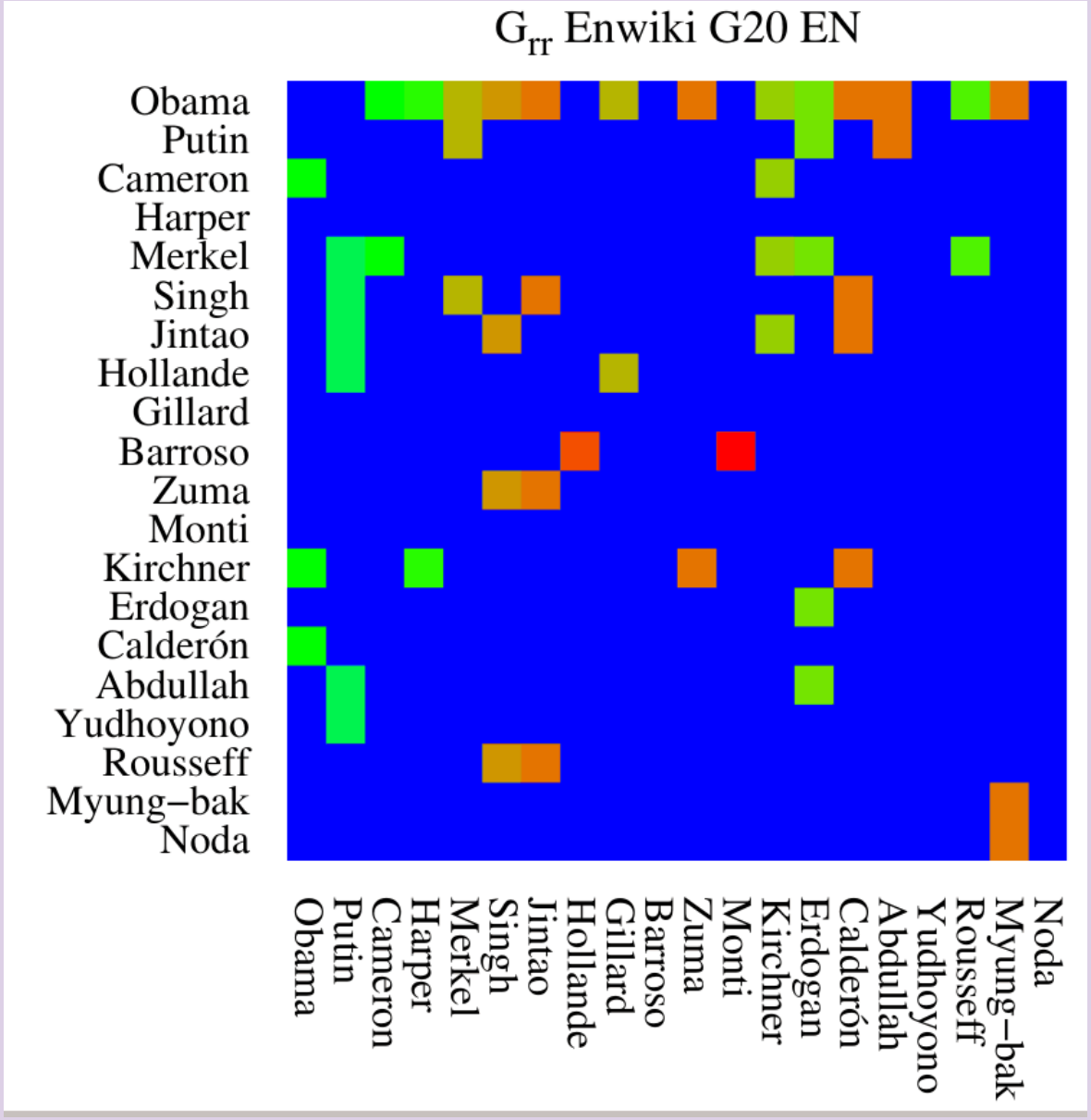
G_{R} G_{pr}



G_{rr} G_{qr}

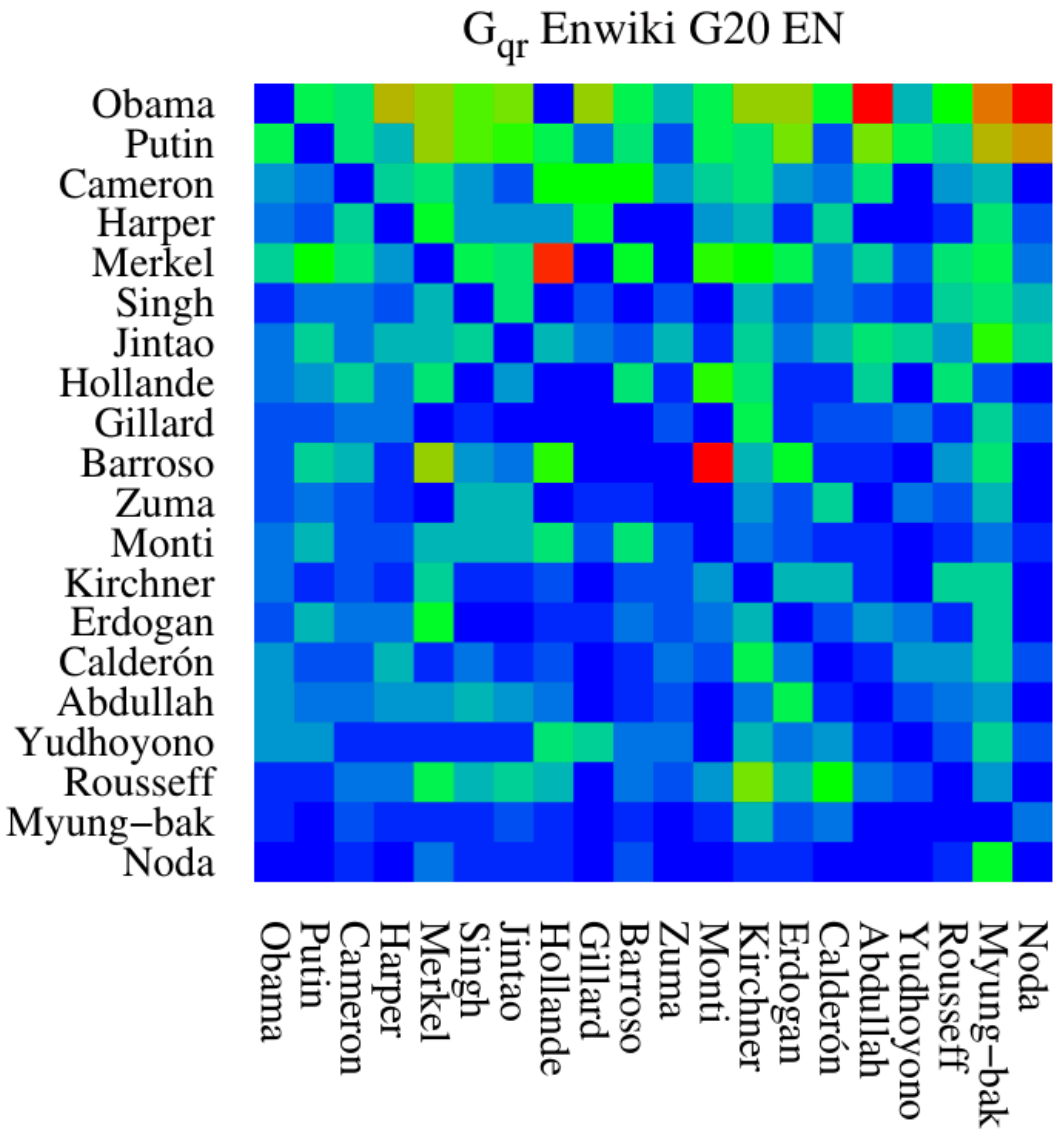
G-reduced: G20 political leaders 2012

G_R example: G20 political leaders 2012 direct links of G_{rr}



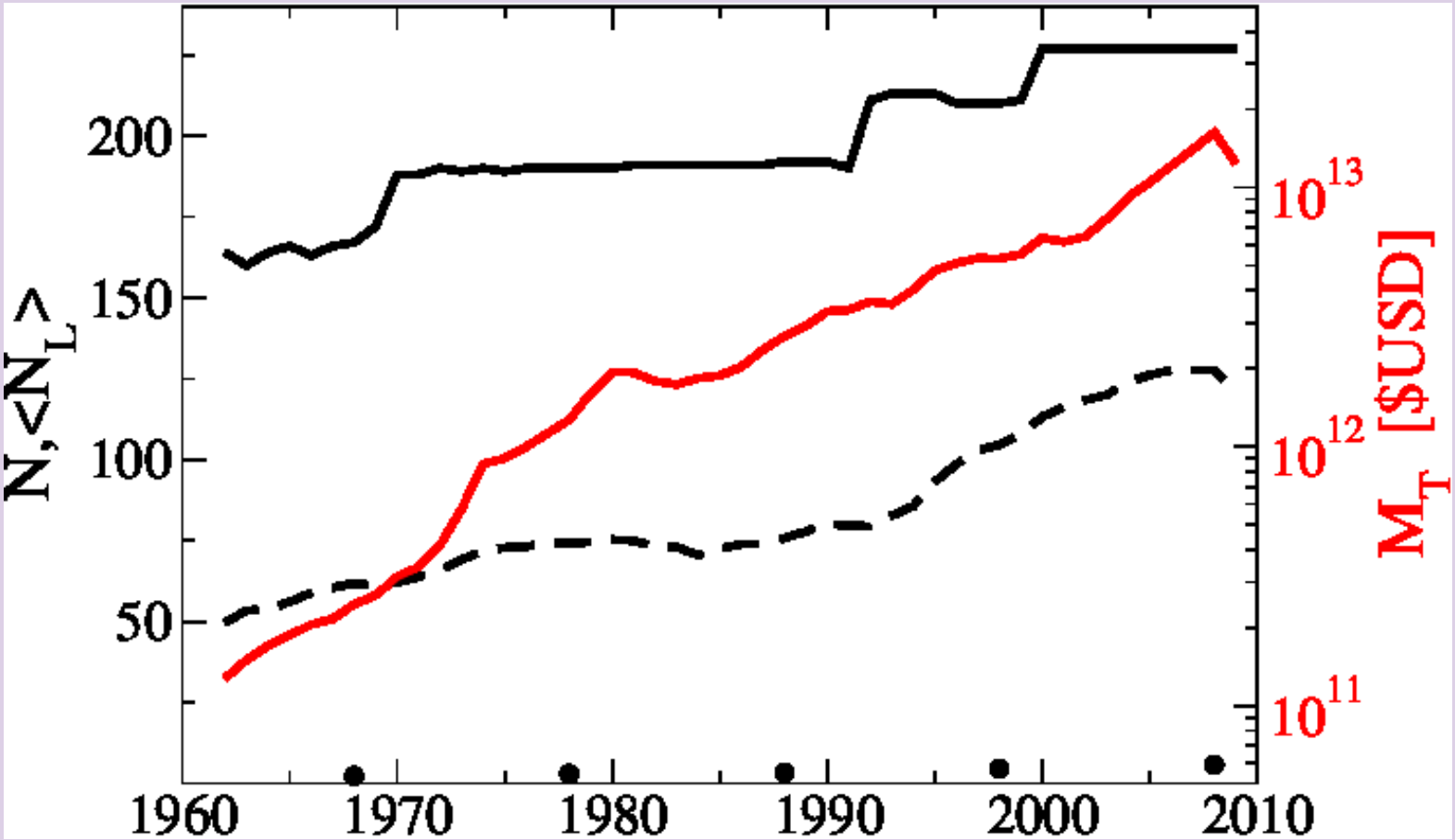
G-reduced: G20 political leaders 2012

G_R example: G20 political leaders 2012 indirect links of G_{qr} (non-diagterms)



G20	EN	Enwiki
Name	Friends	Followers
Obama	Putin	Noda
	Merkel	Abdullah
	Calderón	Myung-bak
Putin	Merkel	Noda
	Obama	Myung-bak
	Barroso	Merkel
Cameron	Putin	Gillard
	Obama	Barroso
	Merkel	Hollande
Harper	Obama	Merkel
	Cameron	Gillard
	Putin	Myung-bak
Merkel	Barroso	Hollande
	Putin	Monti
	Obama	Kirchner

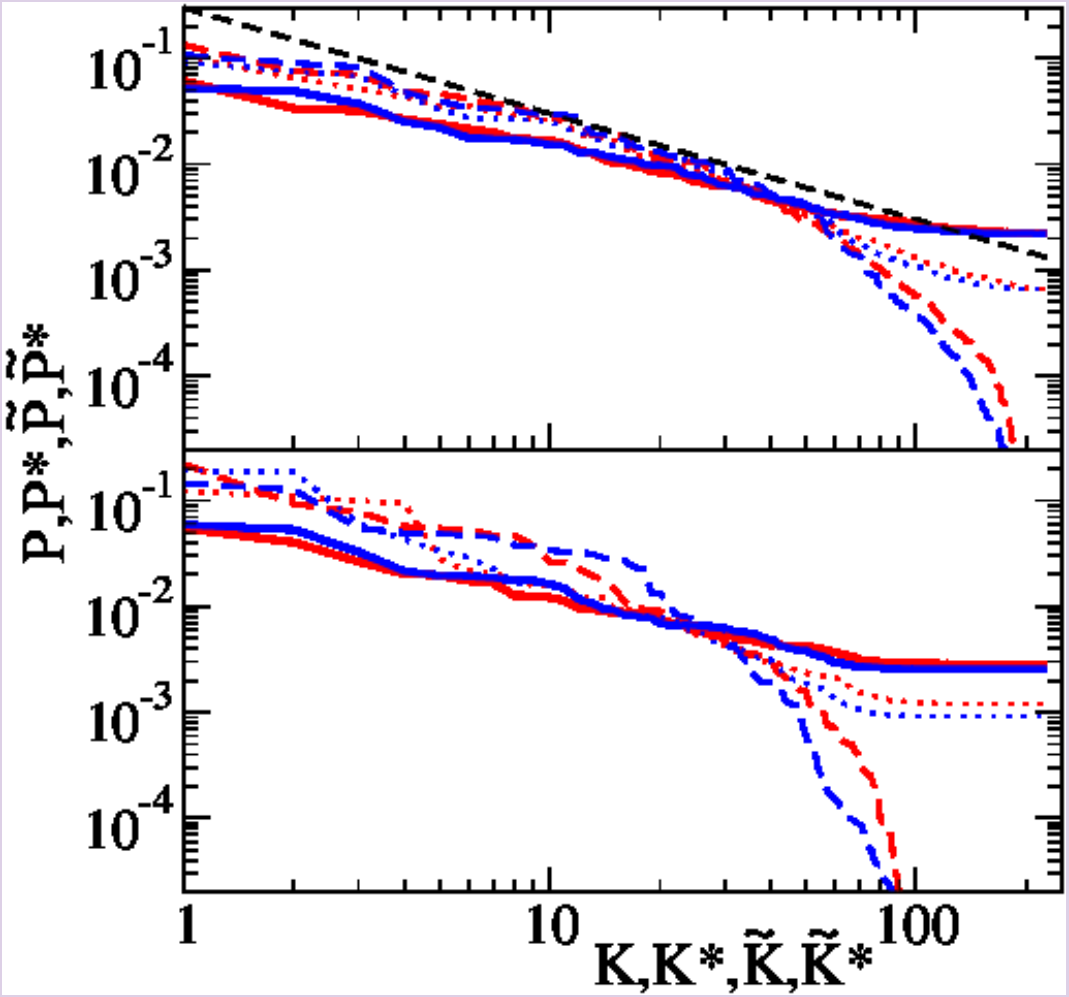
World trade network (WTN) of United Nations COMTRADE 1962-2010



Number of countries (black), links (dashed/points) and mass volume in USD (red)

Leonardo Ermann, DS arxiv:1103.5027 (2011); EPJB (2015)

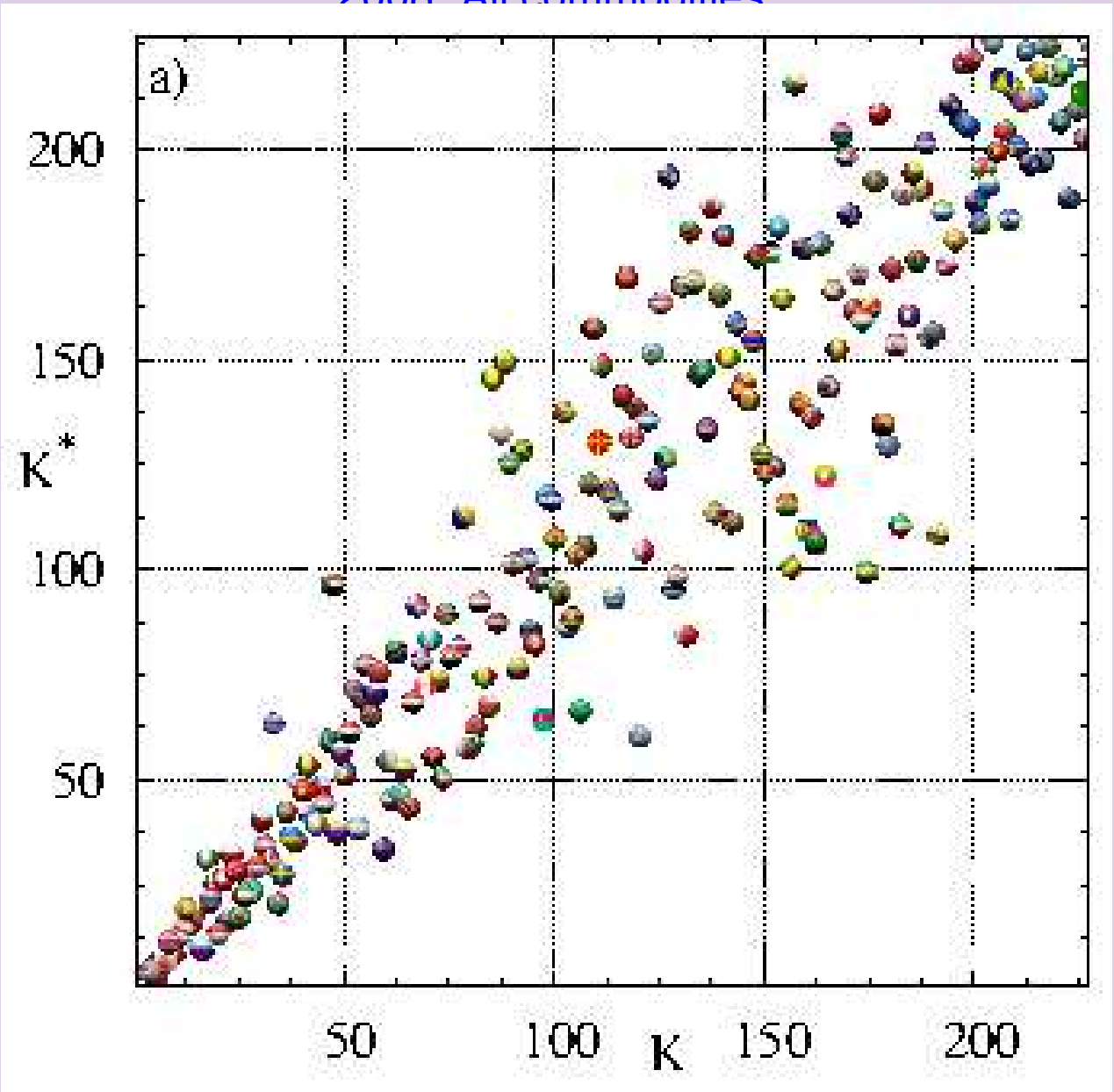
PageRank, CheiRank of World Trade



Year 2008: Probabilities of PageRank $P(K)$ (red), CheiRank $P^*(K^*)$ (blue) for all commodities (top) and crude petroleum (bottom), $\alpha = \mathbf{0.5}; 0.85$ (full/dotted); (dashed curves are for ImportRank, ExportRank); dashed line Zipf law $P \sim 1/K$; 227 countries

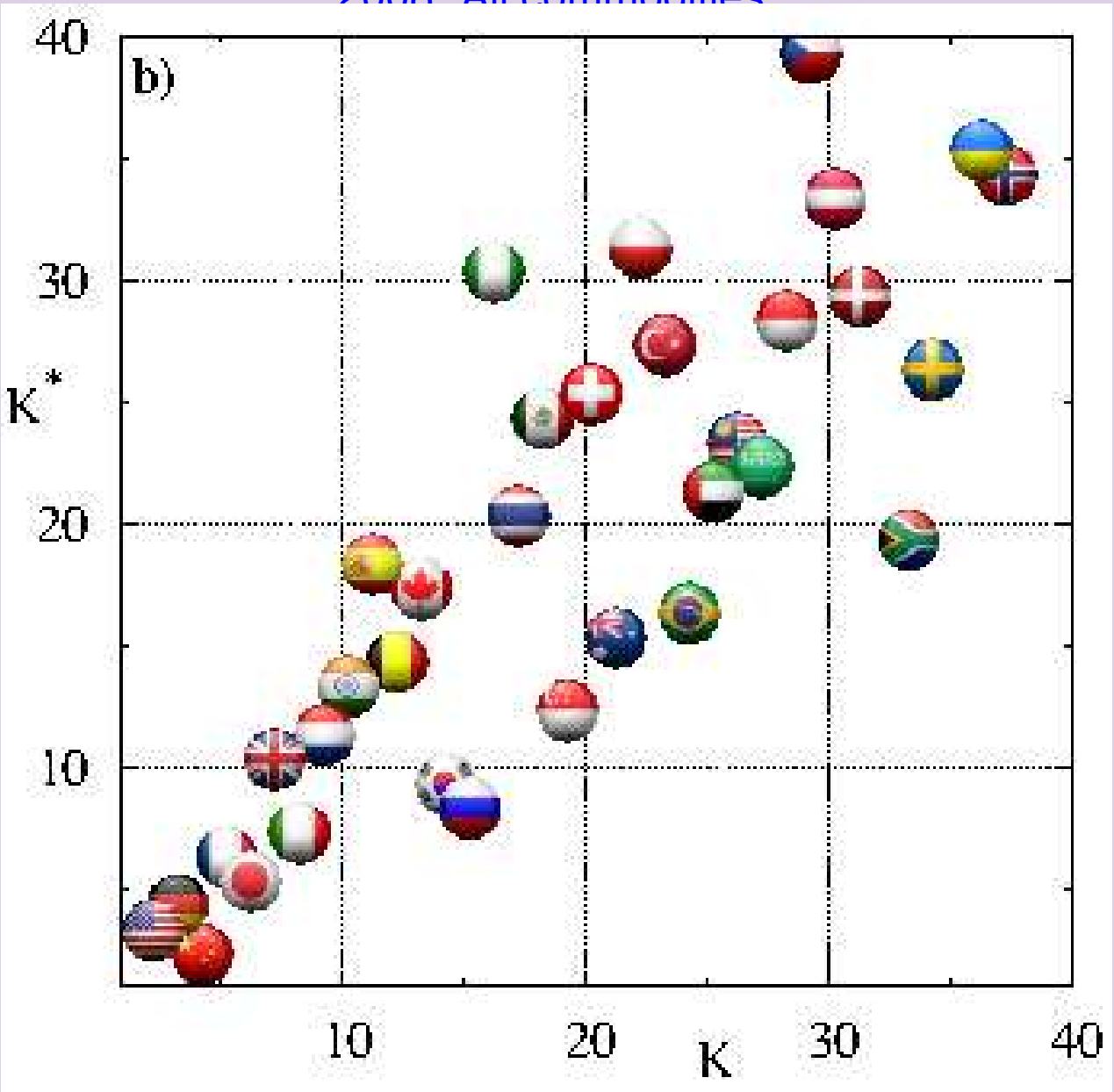
Ranking of World Trade

2008: All commodities



Ranking of World Trade

2008: All commodities



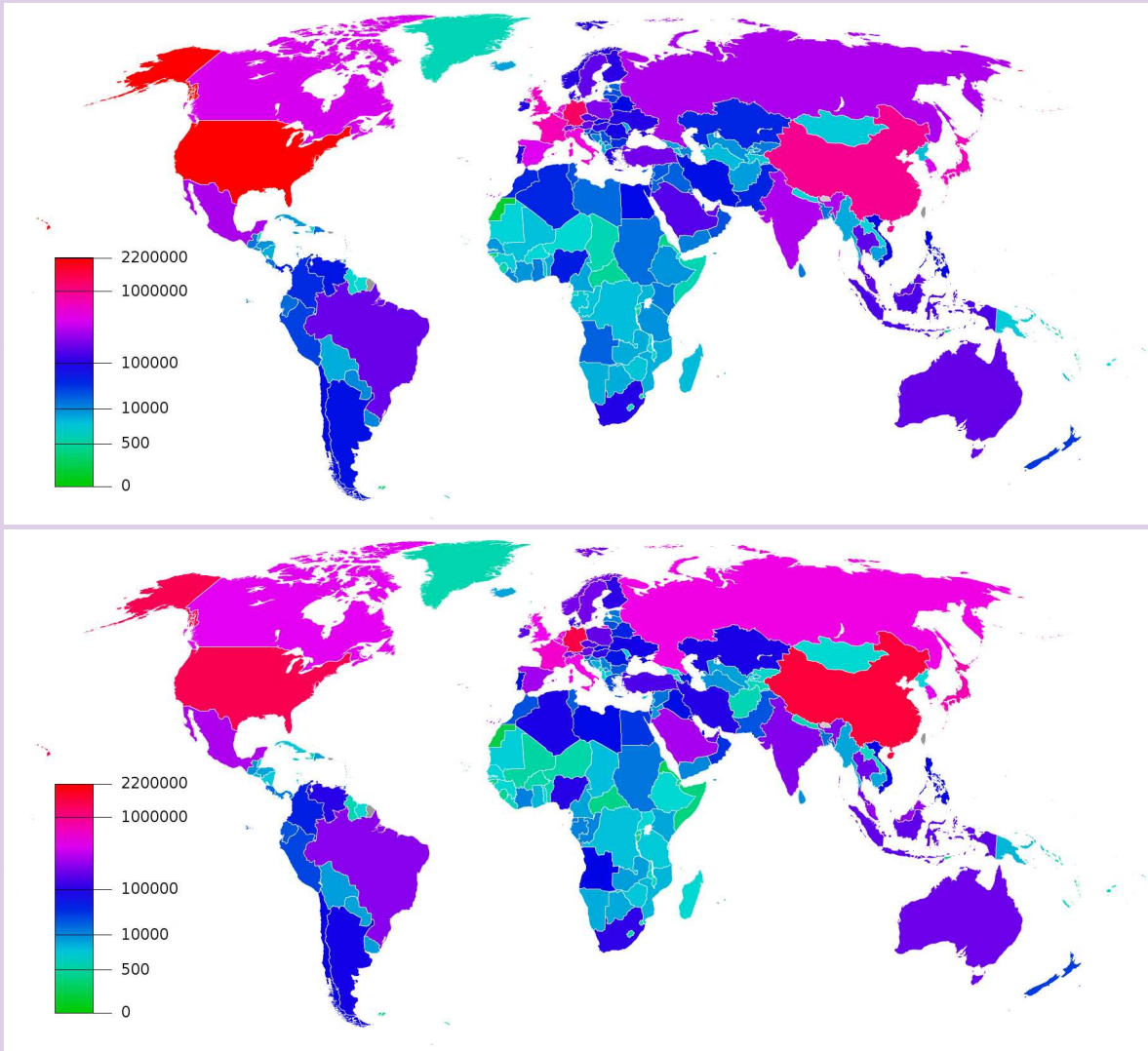
Rank table 2008 (74% of countries of G20)

Table 1. Top 20 ranking for *all commodities* – 2008.

Ran	K	K^*	K_2	$\tilde{K}$	$\tilde{K}^*$
1	USA	China	USA	USA	China
2	Germany	USA	China	Germany	Germany
3	China	Germany	Germany	China	USA
4	France	Japan	Japan	France	Japan
5	Japan	France	France	Japan	France
6	UK	Italy	Italy	UK	Netherlands
7	Italy	Russian Fed.	UK	Netherlands	Italy
8	Netherlands	Rep. of Korea	Netherlands	Italy	Russian Fed.
9	India	UK	India	Belgium	UK
10	Spain	Netherlands	Rep. of Korea	Canada	Belgium
11	Belgium	Singapore	Belgium	Spain	Canada
12	Canada	India	Russian Fed.	Rep. of Korea	Rep. of Korea
13	Rep. of Korea	Belgium	Canada	Russian Fed.	Mexico
14	Russian Fed.	Australia	Spain	Mexico	Saudi Arabia
15	Nigeria	Brazil	Singapore	Singapore	Singapore
16	Thailand	Canada	Thailand	India	Spain
17	Mexico	Spain	Australia	Poland	Malaysia
18	Singapore	South Africa	Brazil	Switzerland	Brazil
19	Switzerland	Thailand	Mexico	Turkey	India
20	Australia	U. Arab Emir.	U. Arab Emir.	Brazil	Switzerland

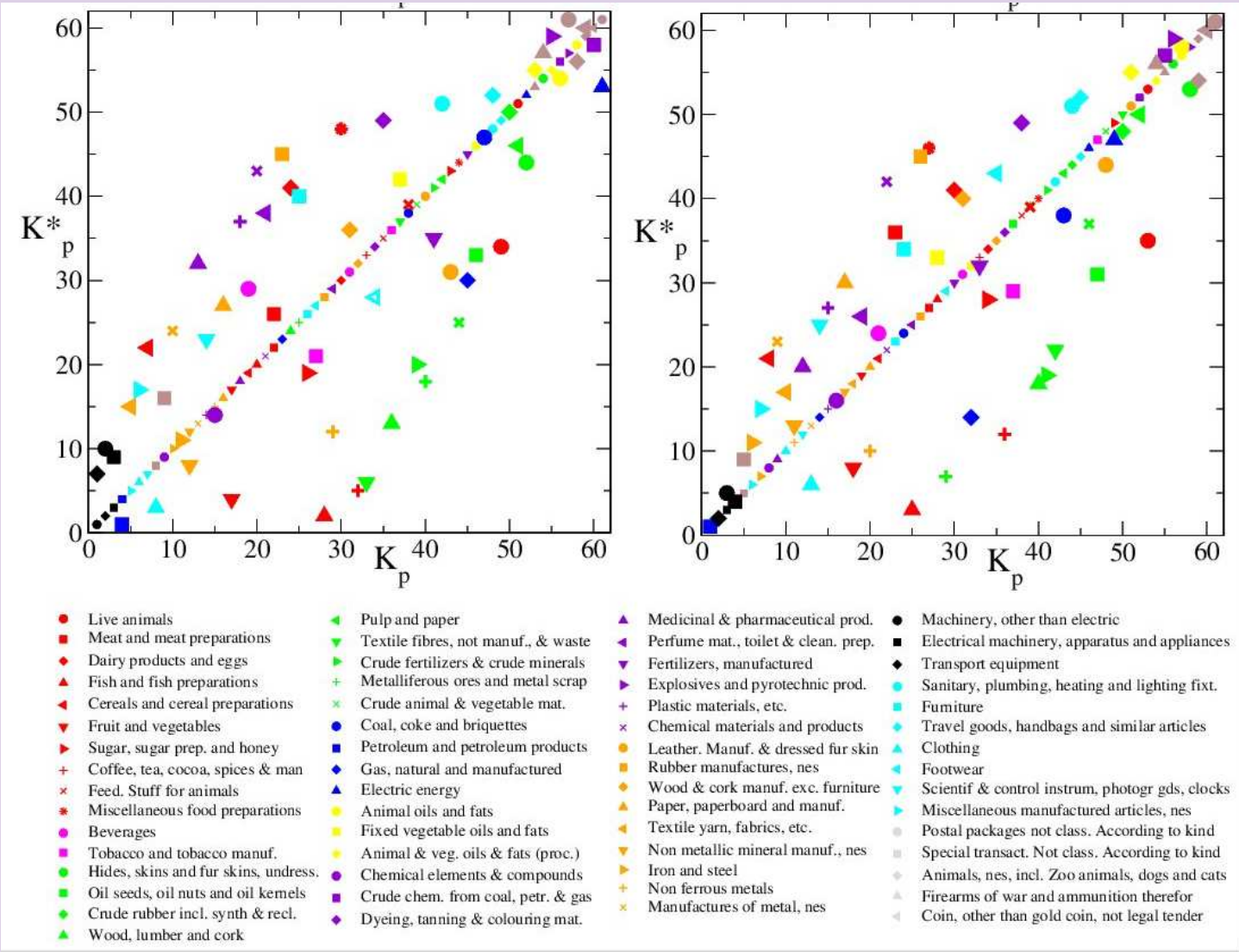
Multiproduct WTN

Example: year 2008, $N_c = 227$, $N_p = 61$, $N = 13847$, import (top) - export (bottom) in millions USD



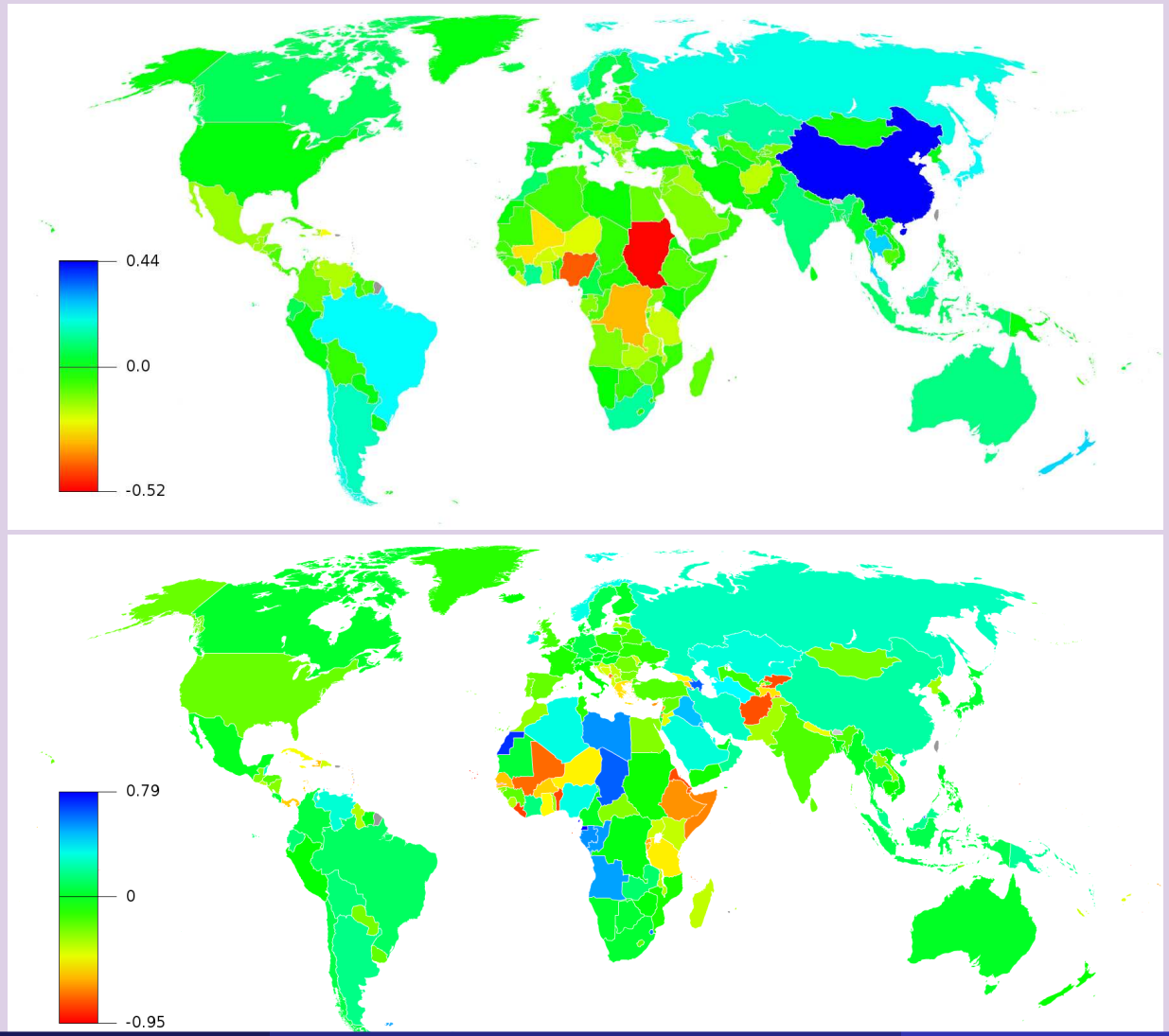
Multiproduct WTN: ranking of products

Democracy in countires, volume fraction in products => personalized vector in *G*. Left: 1993, right: 2008



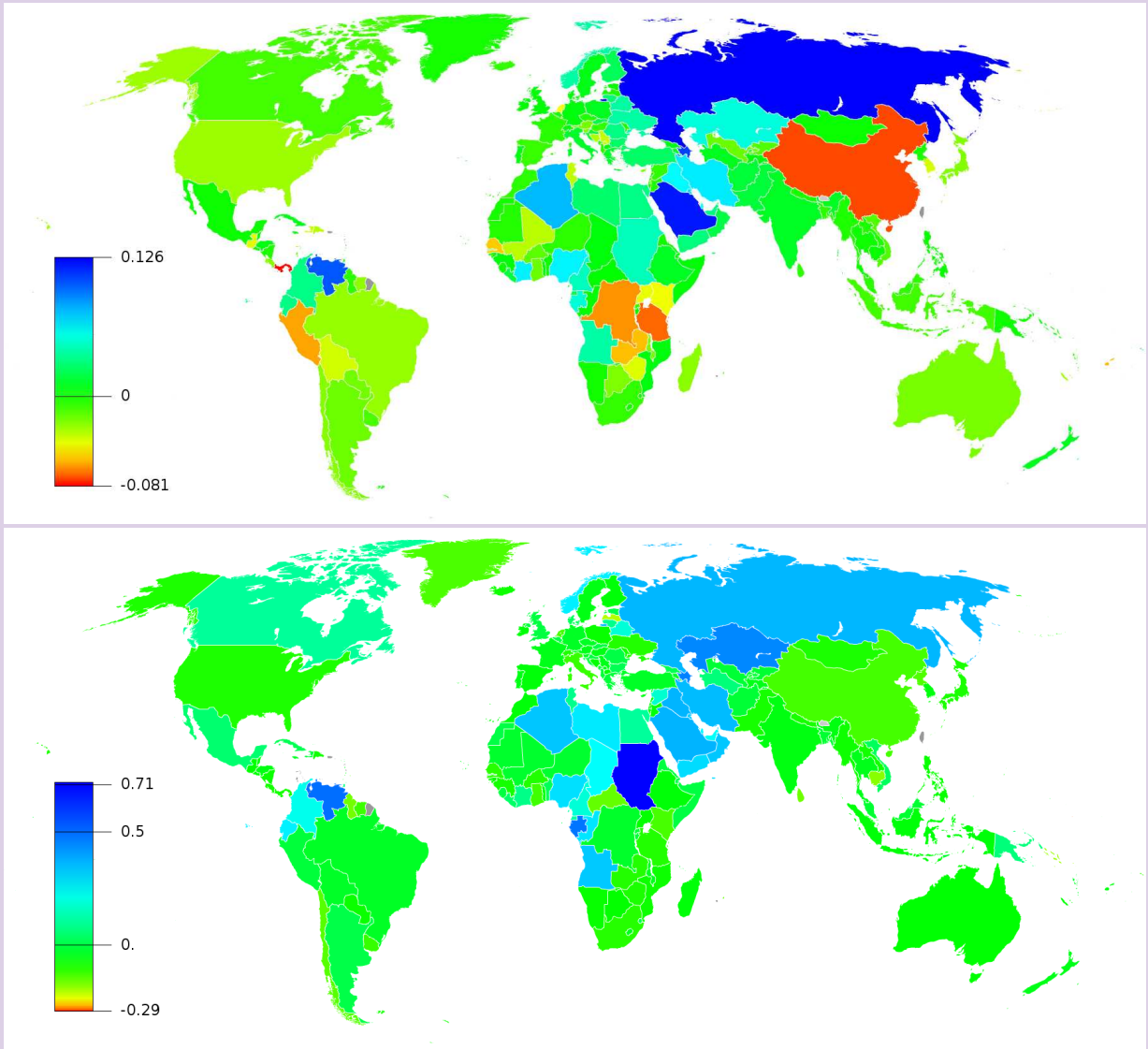
CheiRank-PageRank balance (2008)

$B_c = (P_c^* - P_c)/(P_c^* + P_c)$ (top - CheiRank-PageRank; bottom -Export-Import volume; multiproduct world trade $N_c = 227 countries$, $N_p = 61 products$, $N = 13847$ ==> UN COMTRADE)



Sensitivity to petrolem price (2008)

$B_c = (P_c^* - P_c)/(P_c^* + P_c)$, color => $dB_c/d\delta_{petroleum}$)

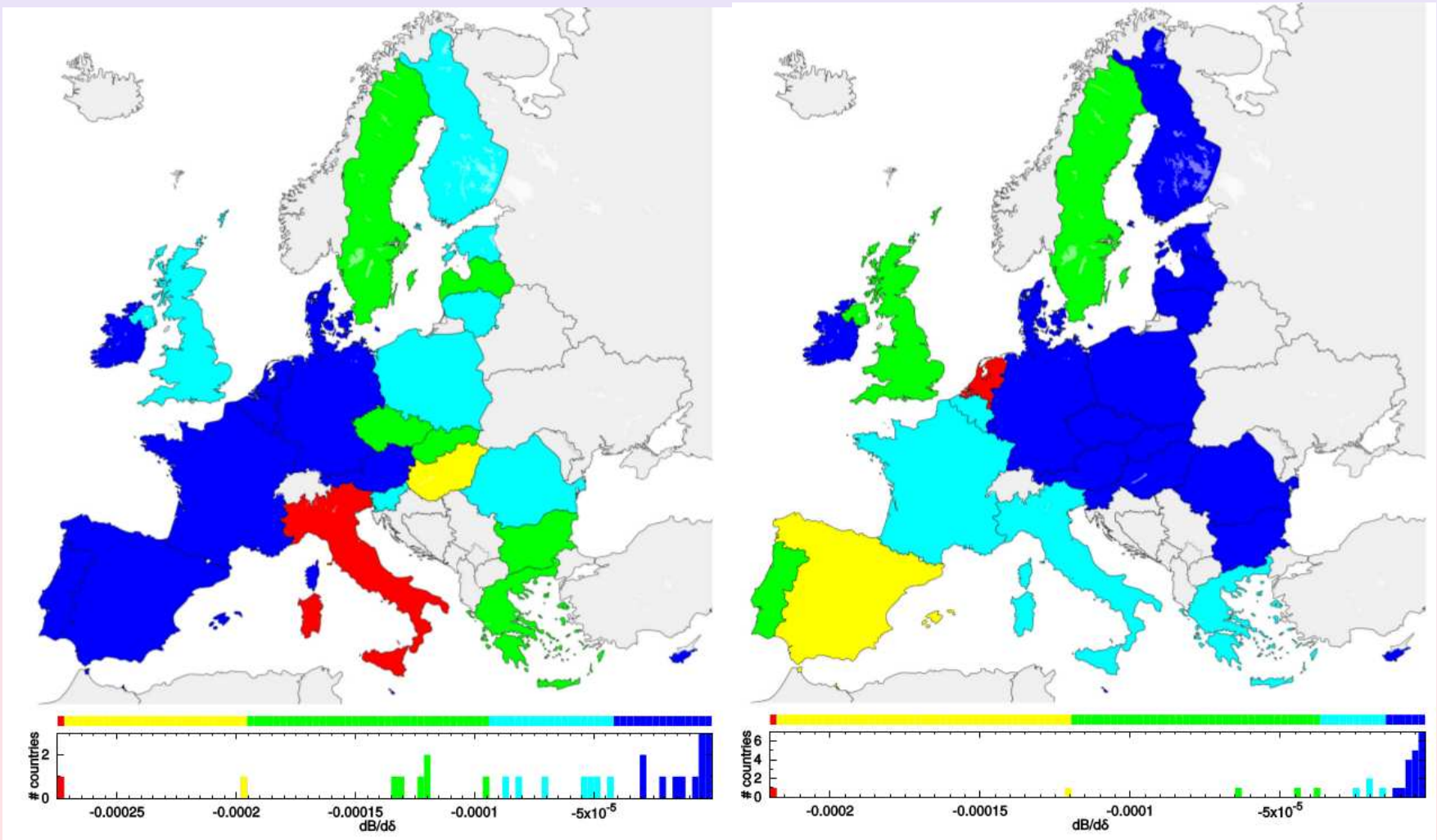


G-reduced for EU sensitivity to RU-US gas

$B_c = (P_c^* - P_c)/(P_c^* + P_c)$, color => $dB_c/d\delta_{gas}$) in 2016

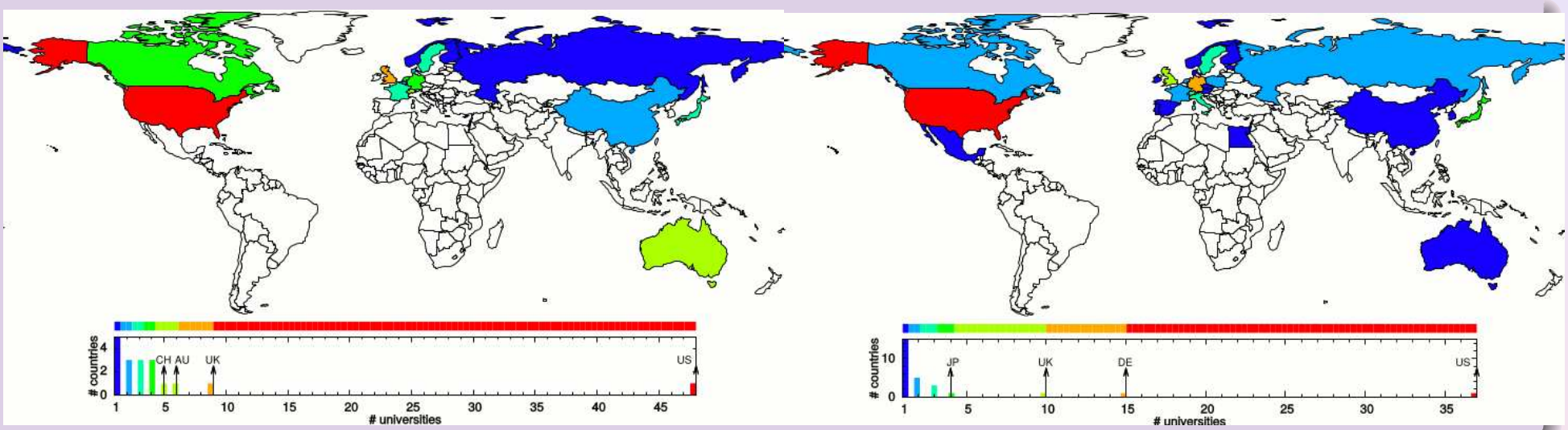
RU/USA -left/right, 27 EU countries of 2008

(Coquide, Ermann, Lages, DS 2019)



Shaghai vs. Wikipedia World University Ranking

Top 100 (Shanghai ARWU2017 - left, wiki WRWU2017 - right), about 20 millions wiki-articles of 24 language editions



ARWU2017: 1.Harvard, 2.Stanford, 3.Cambridge, 4.MIT, 5.Berkeley (...74.Rice; USA: 37)

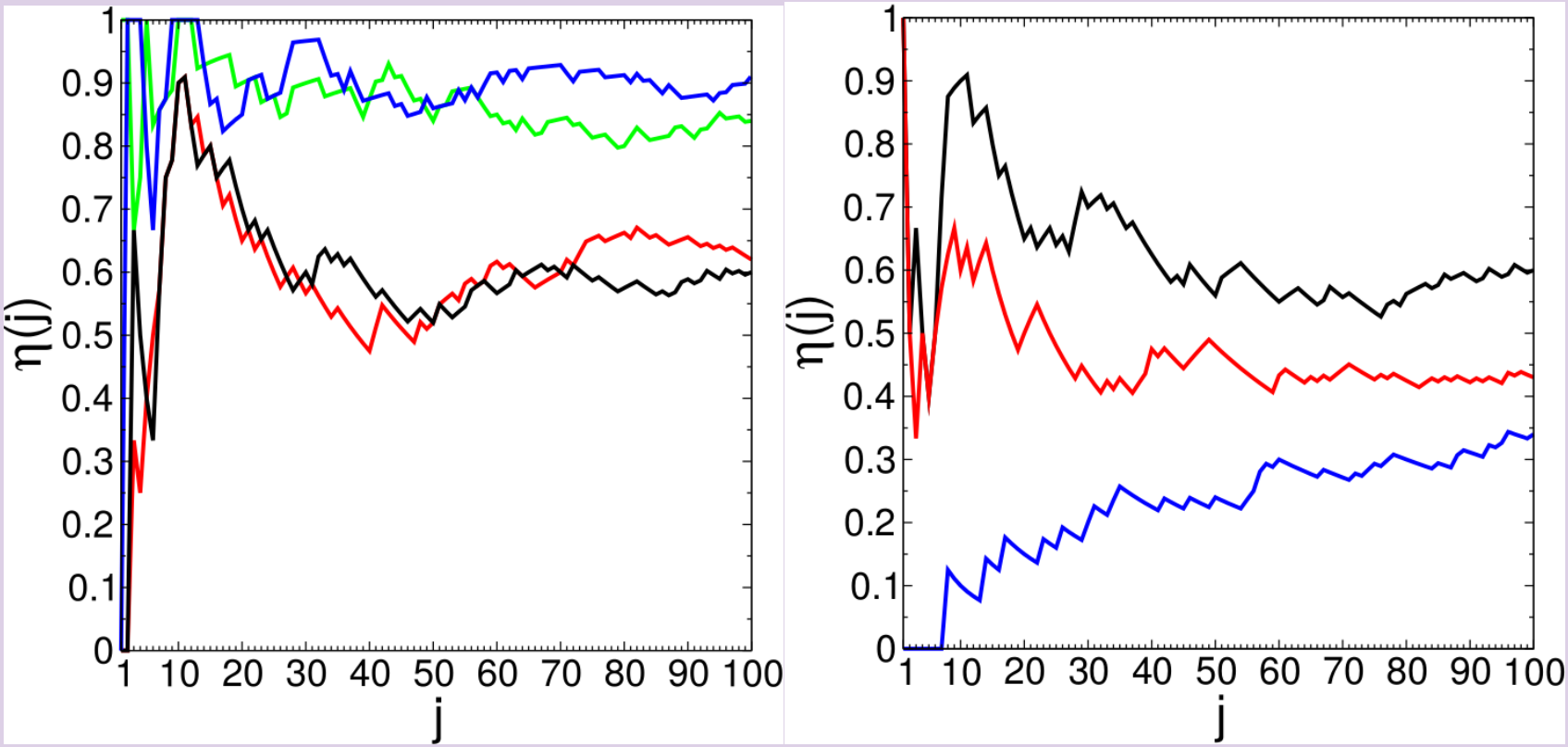
WRWU2017: 1.Oxford, 2.Cambridge, 3.Harvard, 4.Columbia, 5.Yale (...357.Rice; USA: 56)

60percent overlap for top 100

Coquide, Lages, DS (arxiv 2018)

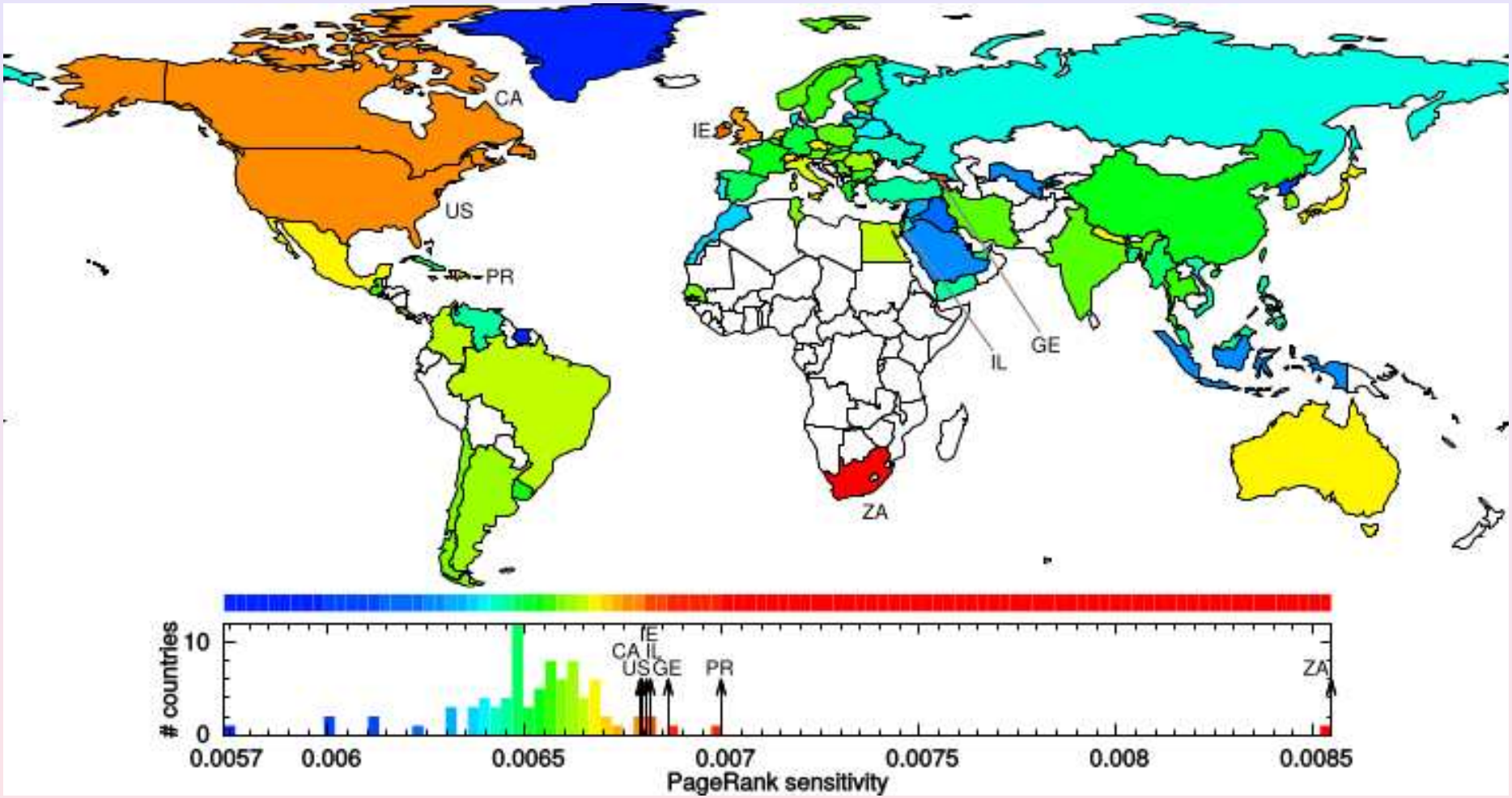
Overlap ARWU, WRWU, EnWiki, FrWiki, DeWiki

Left: overlap ARWU2017 with WRWU2017 (black), ARWU2013 with WRWU2013 (red); ARWU2103 with ARWU2017 (green), WRWU2013 with WRWU2017 (blue); Right: ARWU2017 with EnWiki2017 (black), FrWiki2017 (red), DeWiki2017 (blue)



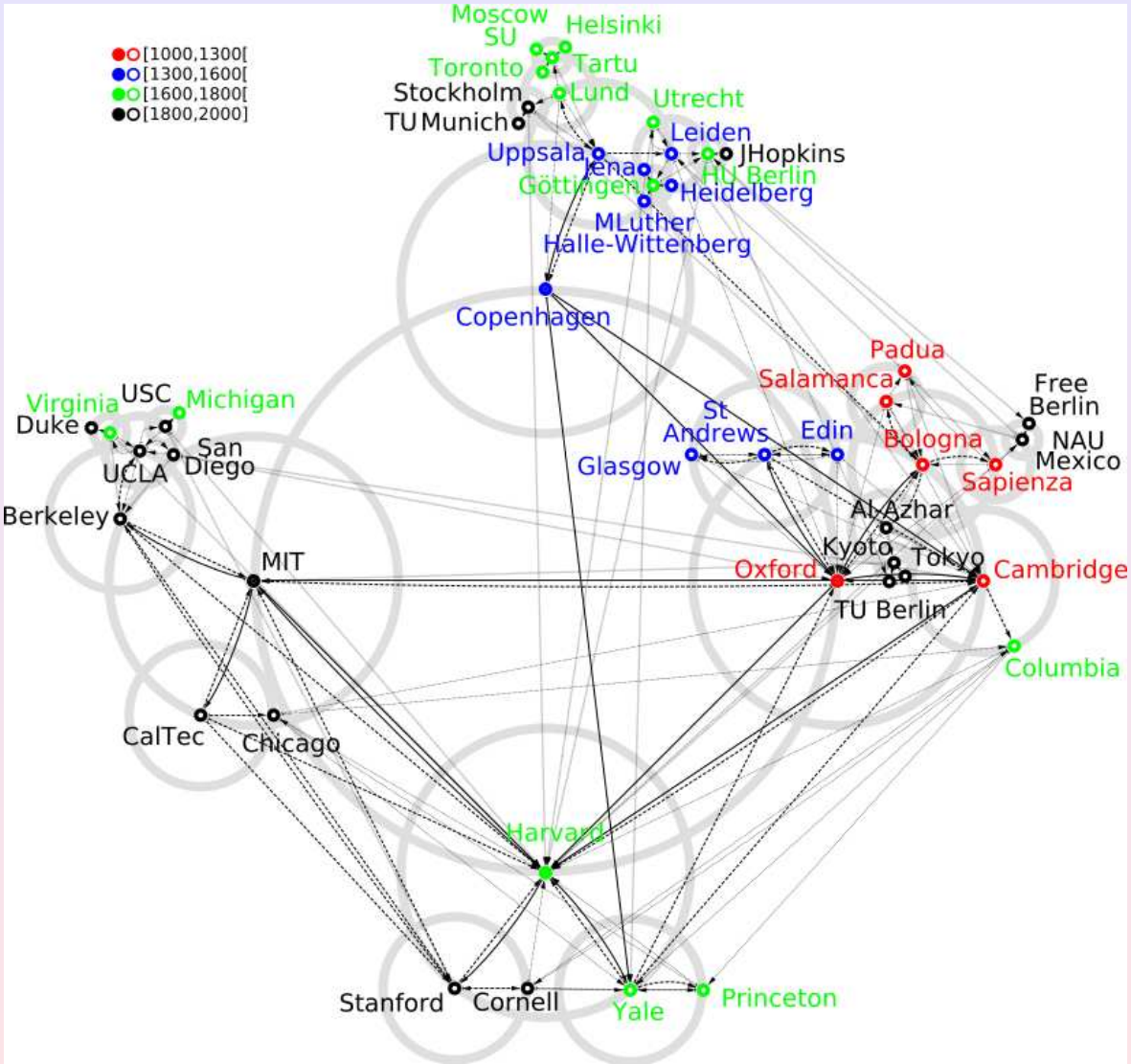
Different cultural views on university ranking
Coquide, Lages, DS (arxiv 2018)

World influence of Harvard University/EnWiki2017



Coquide, Lages, DS (arxiv 2018)

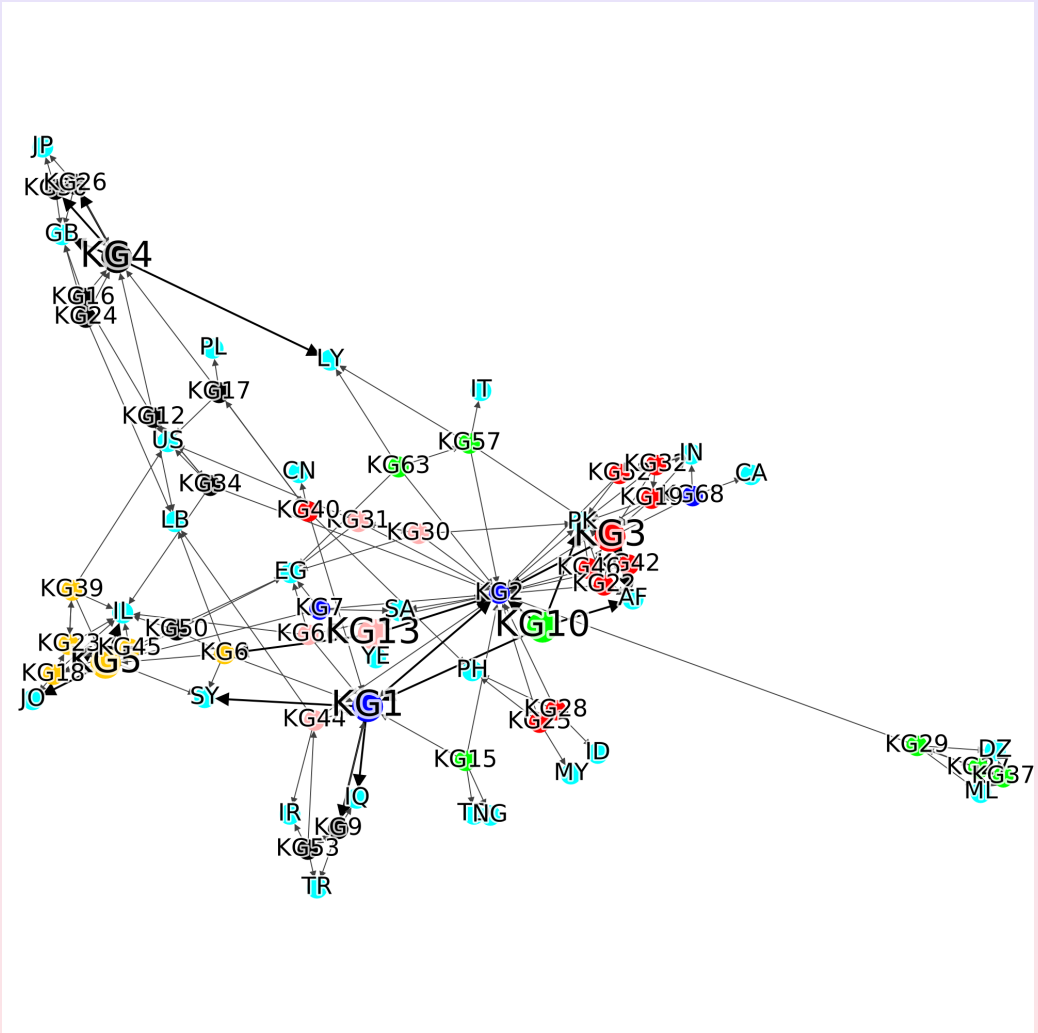
Reduced network of U-friends over centuries



Century reduced friendship network constructed for universities of PageRank top 100 list of WRWU computed from $\bar{G}_{\rm rr} + \bar{G}_{\rm qrnd}$ averaged over 24 Wikipedia editions. Coquide, Lages, DS (arxiv 2018)

Analysis of world terror networks (ENWiki2017)

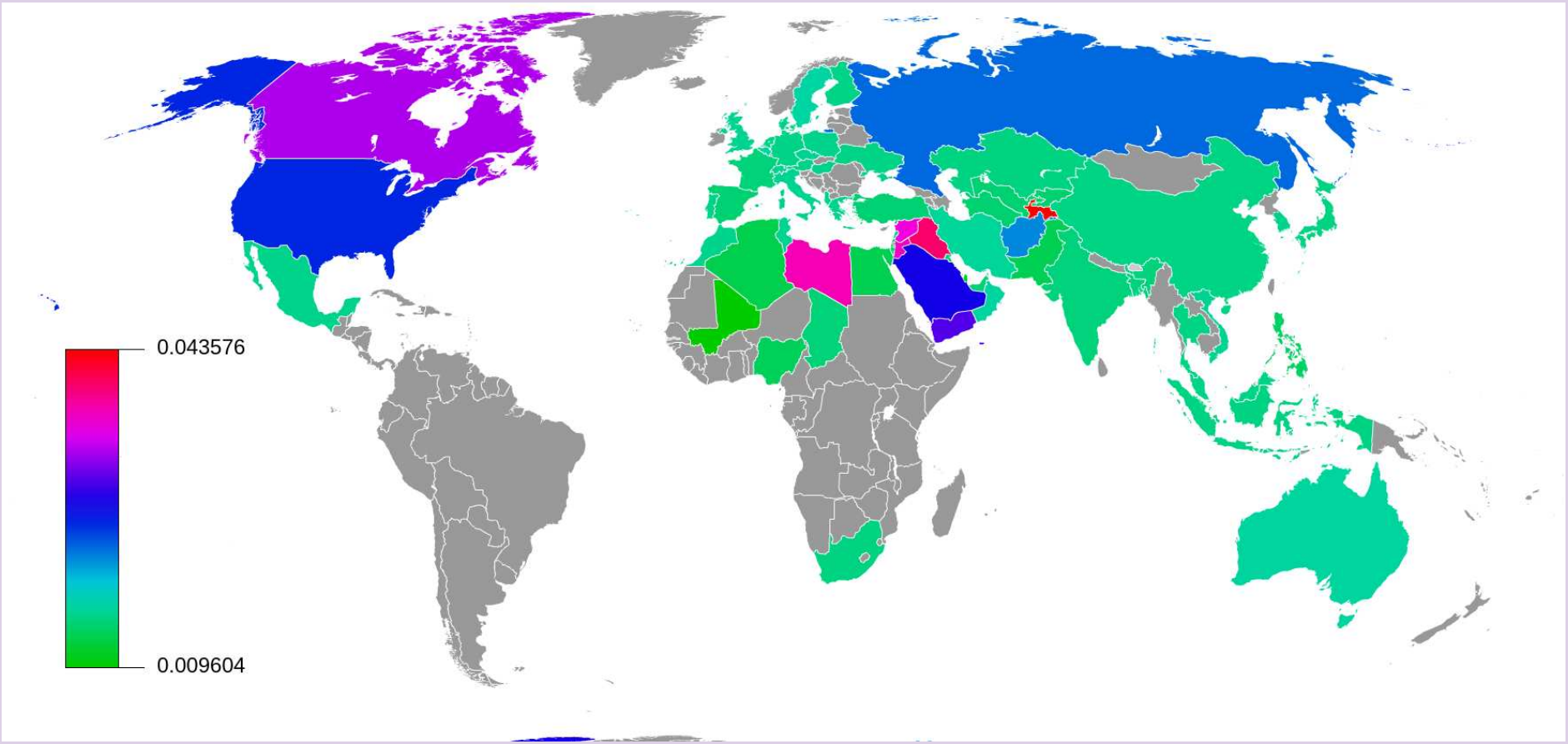
reduced Google matrix for 64 world countries and 95 terrorist groups (from ENWiki2017 with 5.4 millions articles)



Friendship network structure from $G_{qr} + G_{rr}$ with the top terrorist groups (marked by their respective colors) and countries (marked by cyan color); it is shown with 2 friends for top terrorist groups of each category and top friend 2 countries for each group. El Zant, Frahm, Jaffres-Runser, DS (EPJB 2018)

Analysis of world terror networks (ENWiki2017)

reduced Google matrix for 64 world countries and 95 terrorist groups (from ENWiki2017 with 5.4 millions articles)

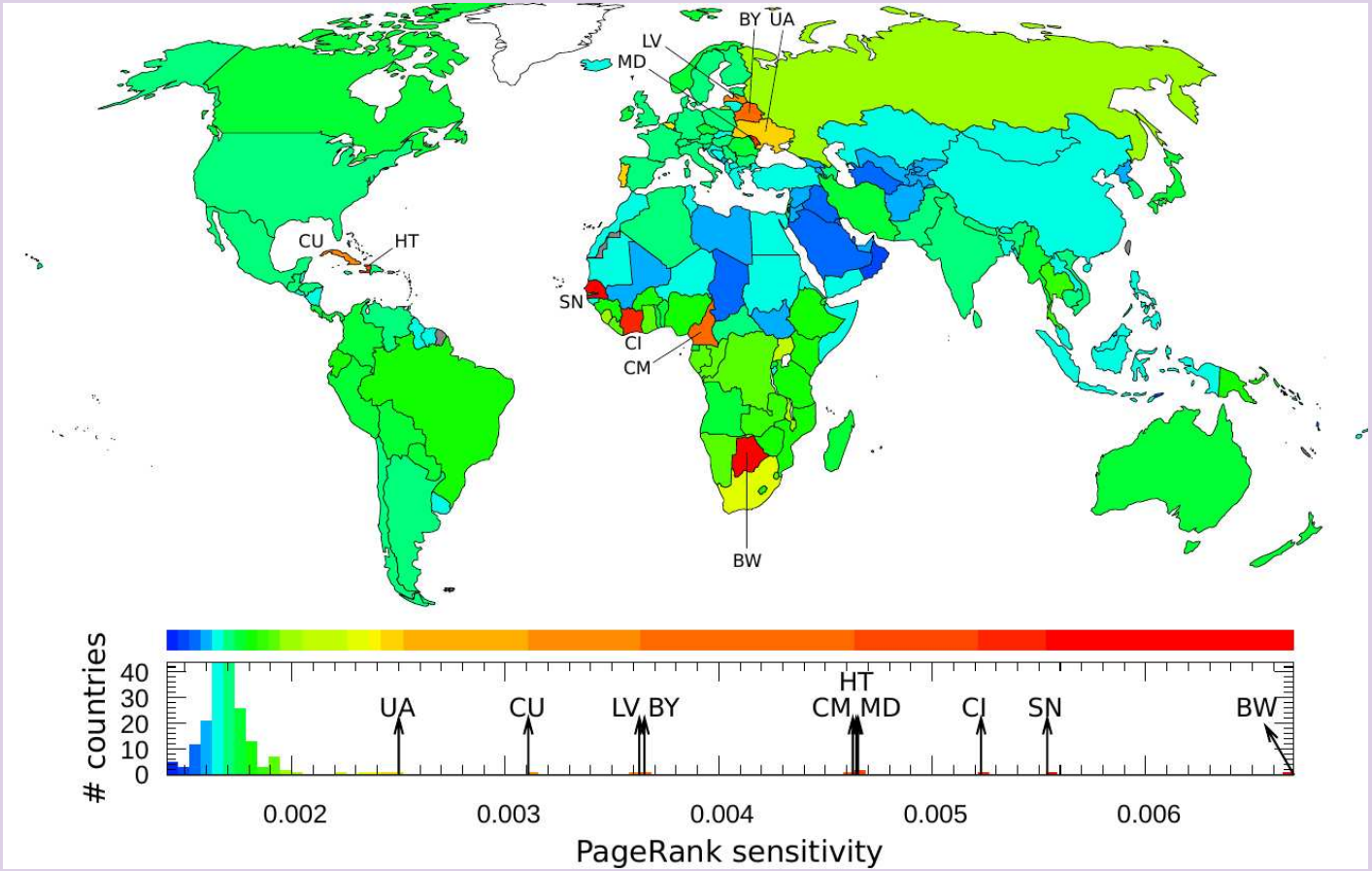


PageRang sensitivity D to weight link variation ISIS - country $(dP_c/(P_c d\delta))$

El Zant, Frahm, Jaffres-Runser, DS (EPJB 2018)

HIV(AIDS) influence of world countries (ENWiki2017)

reduced Google matrix for 195 world countries and 230 infectious diseases (from ENWiki2017 with 5.4 millions articles)

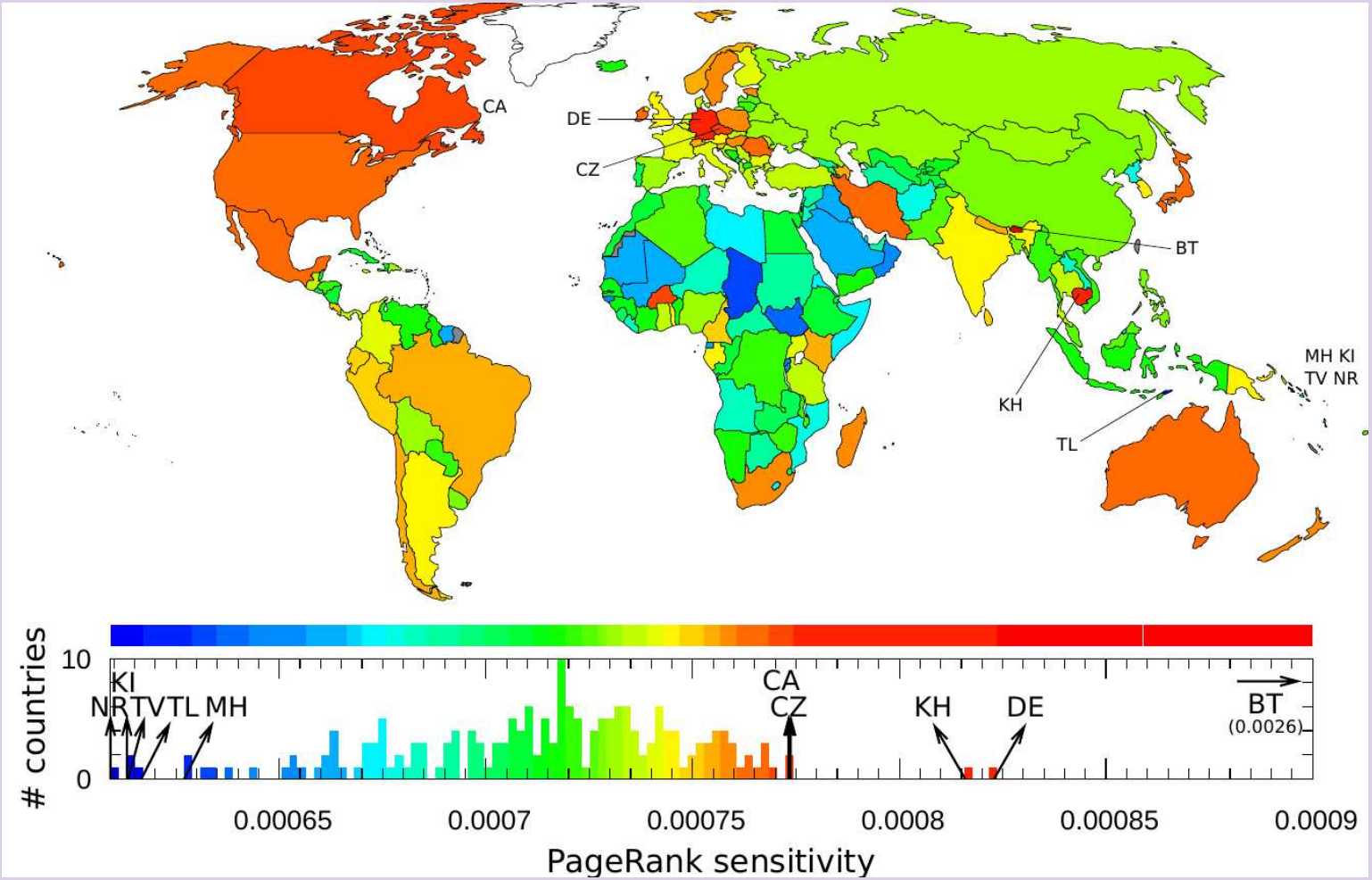


PageRang sensitivity D to weight link variation HIV-country $(dP_c/(P_c d\delta))$

Rollin, Lages, DS (bioRxiv 2018)

World impact of lung cancer (ENWiki2017)

reduced Google matrix for 195 world countries, 37 cancers and 203 drugs (from ENWiki2017 with 5.4 millions articles)



PageRang sensitivity D to weight link variation lung cancer to country $(dP_c/(P_c d\delta))$ (Rollin, Lages, DS (in progress 2018))

Towards bank financial network ranking

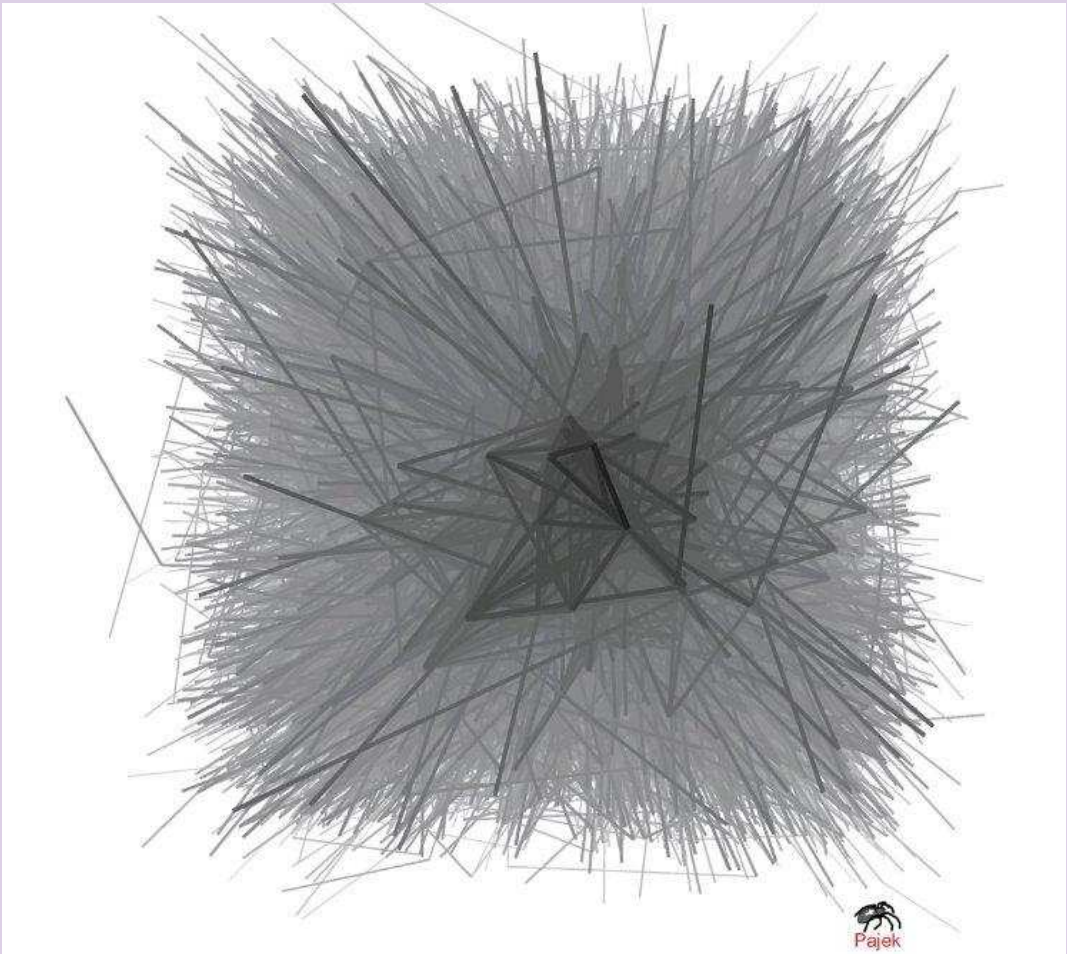


Fig. 1. Fedwire interbank payment network. First day of Sample. 6600 nodes and over 70,000 undirected links [39].

K.Soramäki *et al.*, *The topology of interbank payment flows*, Physica A **379**, 317 (2007) (only about 6000 nodes for all Federal Reserve USA)

2000 bank units for all Germany

Possible ALTRAN+Quantware network projects

* World network of flights:
network of flights between world cities (undirected ?);
weighted with aviation fuel of flight (directed network); $N \sim 10^4$
(proposal of Nicolas Martin)

* Industrial networks of control:
e.g. interactions between sensors ?

* Reduced Google matrix for Twitter network:
determine leading nodes, keywords and their interactions ...

* Financial flows and bank transactions: ?

* Network of protein-protein interactions in biological networks:
SIGNOR https://signor.uniroma2.it/ open public
REACTOME https://reactome.org/ open public
METACORE https://clarivate.com/products/metacore/ private
(request of millions of USD for the whole network of about 20000 nodes)

* * *

see available tools at Ermann, Frahm, DS Rev Mod Phys (2015)
http://www.quantware.ups-tlse.fr/dima/subjgoogle.html

Further applications of Markov chains and Google matrix for ALTRAN? Artificial intelligence at Toulouse?

Google matrix: fundamentals, applications and beyond, IHES workshop 15-18 Oct 2018 (see www.ihes.fr)