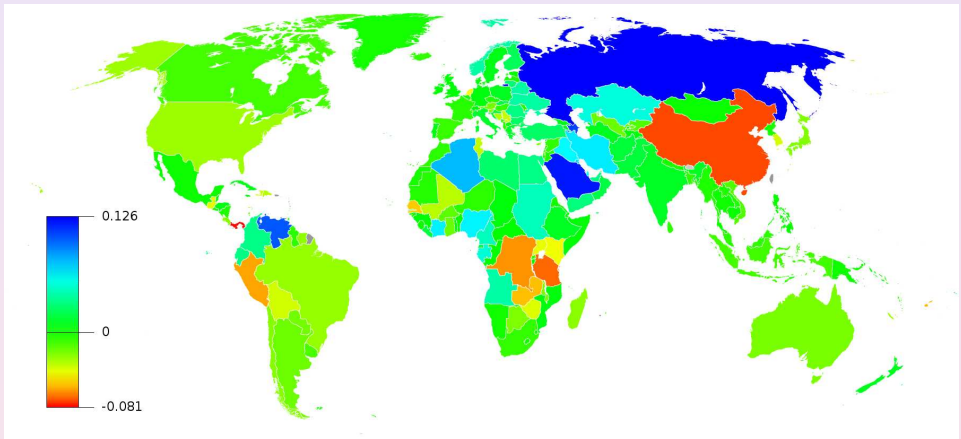
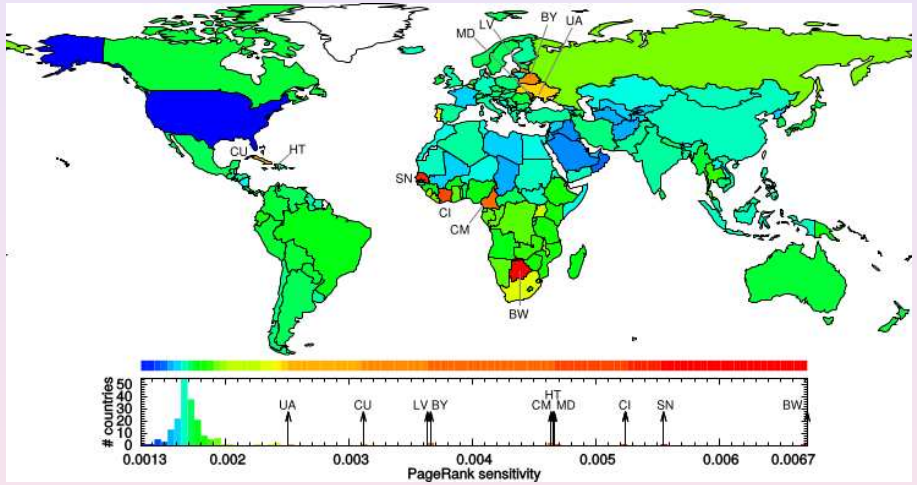
Maps of influence and interactions of infectious diseases from Wikipedia networks

Dima Shepelyansky (CNRS, Toulouse)
www.quantware.ups-tlse.fr/dima
with J.Lages and G.Rollin (U Besancon)

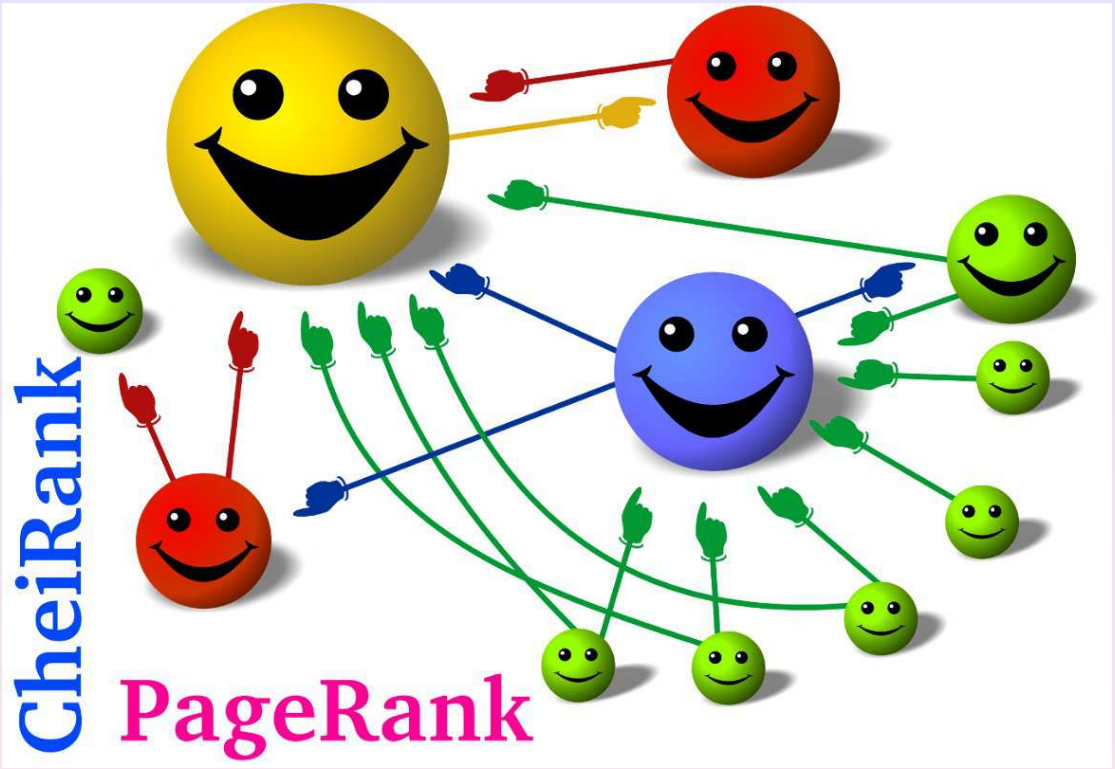


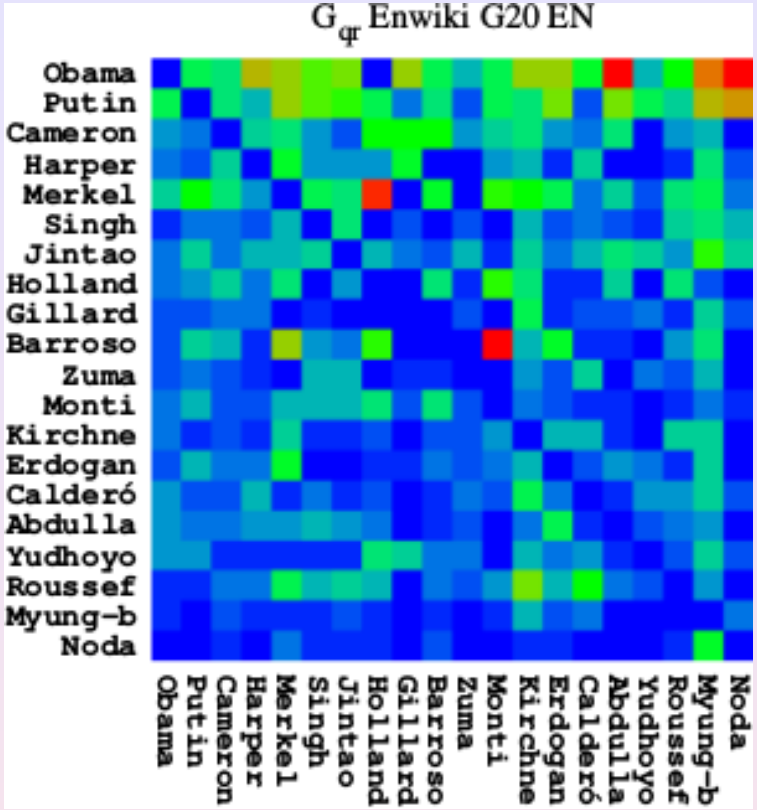


* Markov (1906) $\rightarrow$ Brin and Page (1998)
* reduced Google matrix of directed networks (brief introduction)
* Applications: multiproduct world trade network (UN COMTRADE), Wikipedia Ranking of World Universitis (WRWU), protein-protein interactions, ...
* diseases and drugs influence from English Wikipedia 2017 (5.4 millions articles)

Support: EC FET Open NADINE, APLIGOOGLE (CNRS) 2012-2017, LABEX NEXT THETRACOM (2017-2019) + Rev. Mod. Phys. 87, 1261 (2015)

(1906) Markov vs Wigner (1955)





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

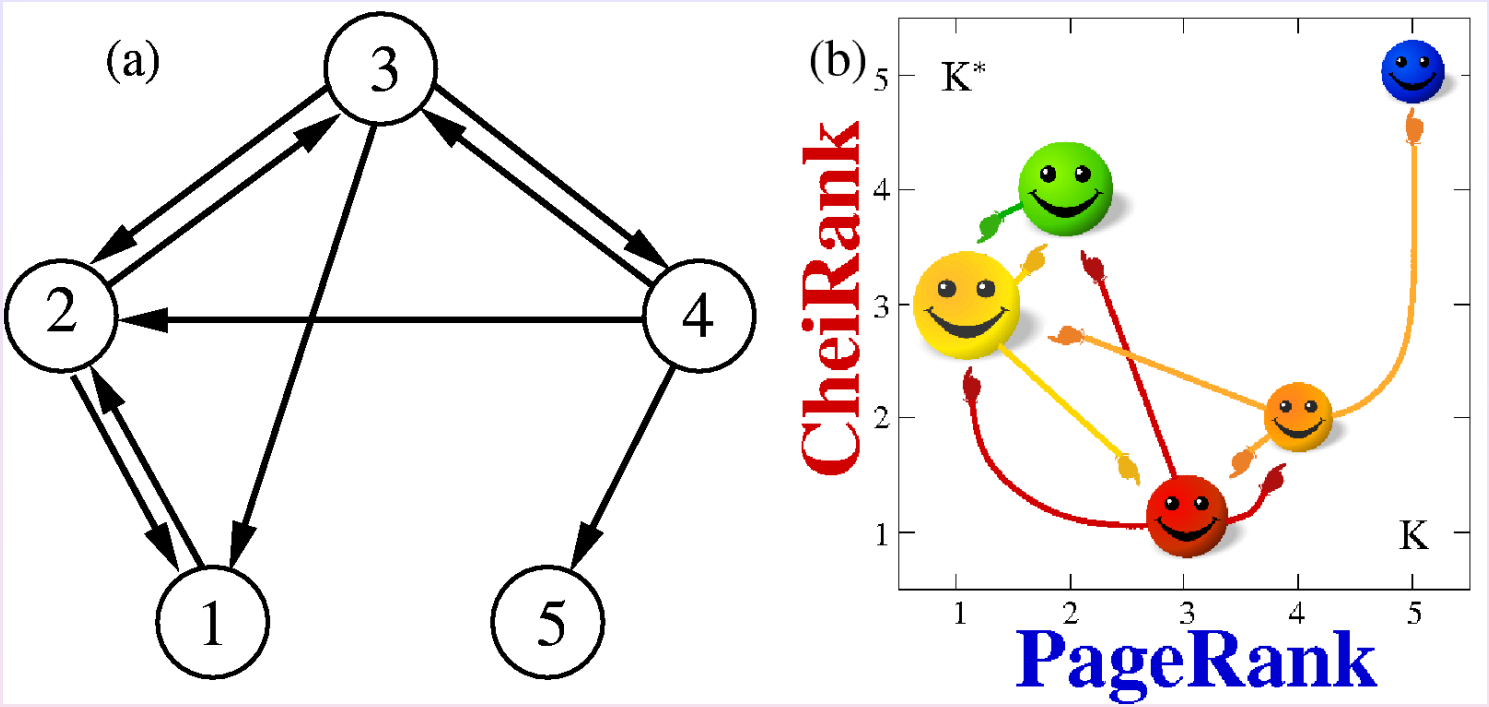
1991: WWW, small world social networks $\rightarrow$ Markov (1906) $\rightarrow$ Google matrix

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) ...

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

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

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

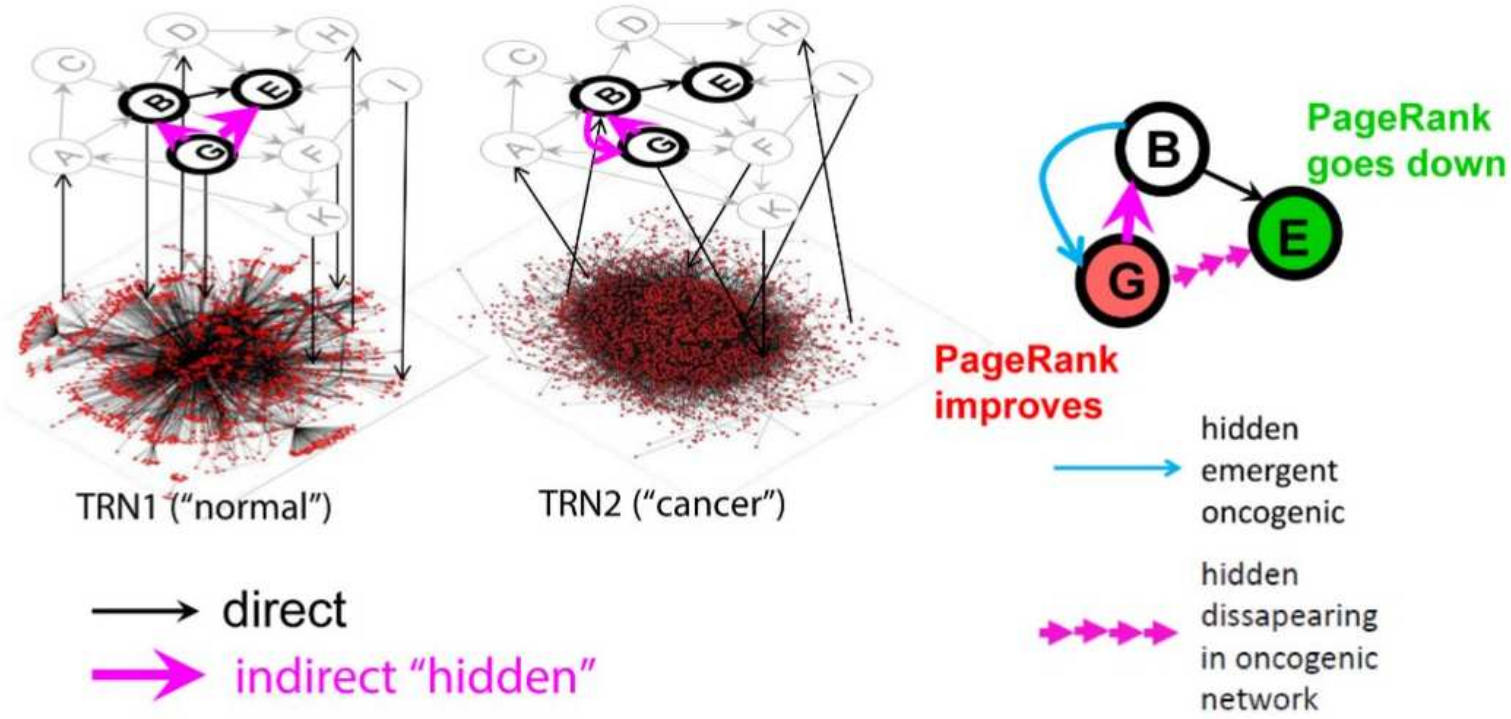


Fig 1. Using reduced Google matrix approach for inferring hidden causal relations in signaling pathways. Here the structure of the context-dependent global regulatory network is symbolically shown as consisting of two layers: the upper (nodes A-K) is the global signaling network whose structure does 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)

PageRank of Infectious diseases - EnWiki2017

Type of infectious diseases : bacterial, viral, parasitic, fongic, prionic, multiple origin, other

Rank	Disease
1	Tuberculosis
2	HIV/AIDS
3	Malaria
4	Pneumonia
5	Smallpox
6	Cholera
7	Influenza
8	Measles
9	Typhoid fever
10	Syphilis
11	Yellow fever
12	Bubonic plague
13	Poliomyelitis
14	Leprosy
15	Sepsis
16	Meningitis
17	Plague
18	Typhus
19	Diphtheria
20	Severe acute respiratory syndrome

230 diseases, 195 countries (5.4/122 million articles/links)

Network structure, PageRank-CheiRank plane

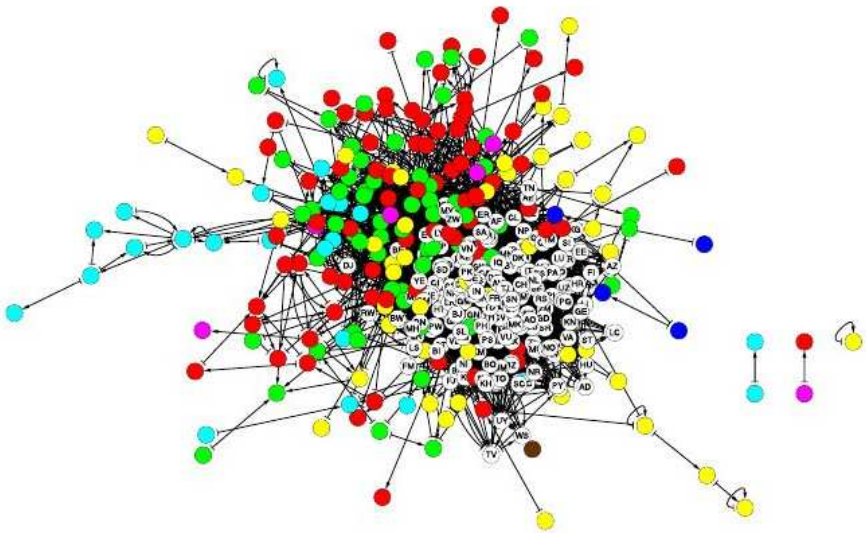
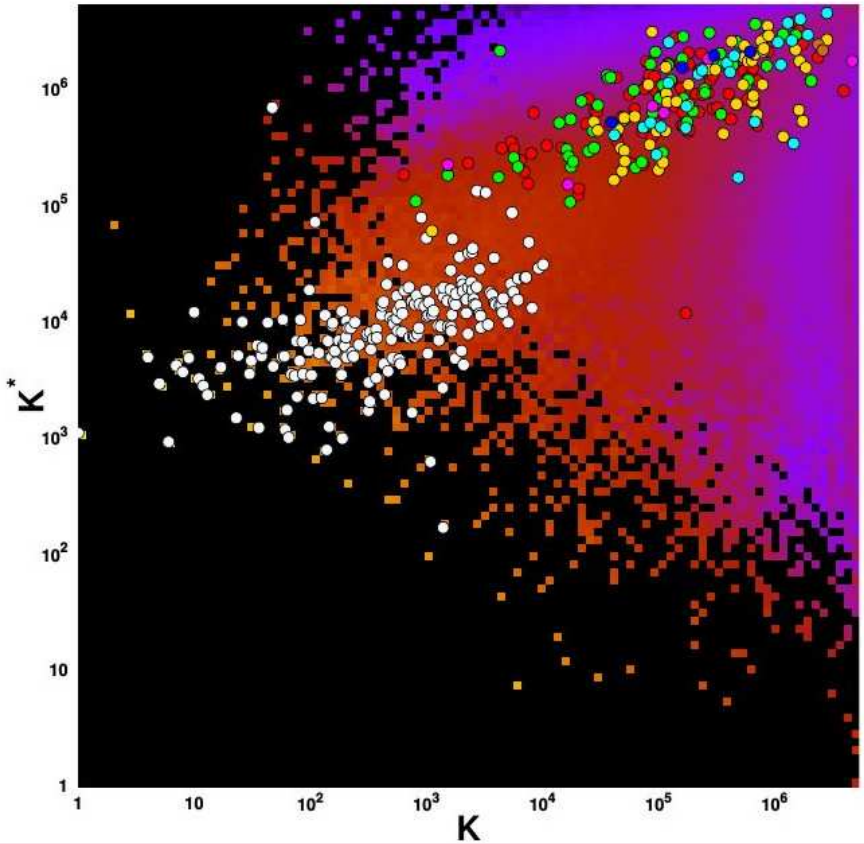
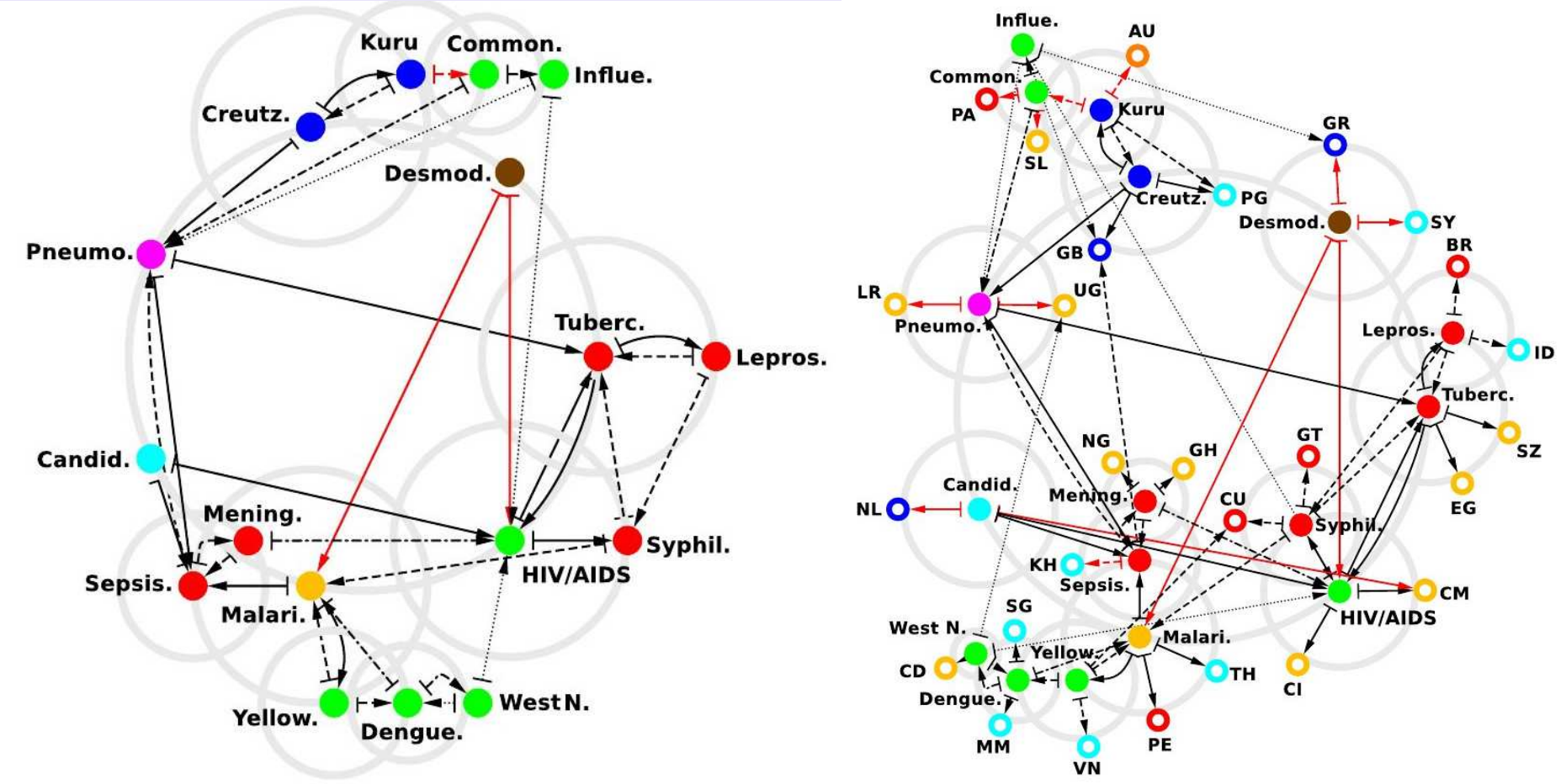


FIGURE 1. Subnetwork of the 425 articles devoted to countries and infectious diseases in 2017 English Wikipedia. The bulk of Wikipedia articles is not shown. Articles devoted to countries are presented by empty nodes with country codes (see Tab. 3). Articles devoted to infectious diseases are presented by colored nodes with the following color code: **bacterial diseases**, **viral diseases**, **parasitic diseases**, **fongic diseases**, **prionic diseases**, **diseases with multiple origins**, and **other kind of diseases**.



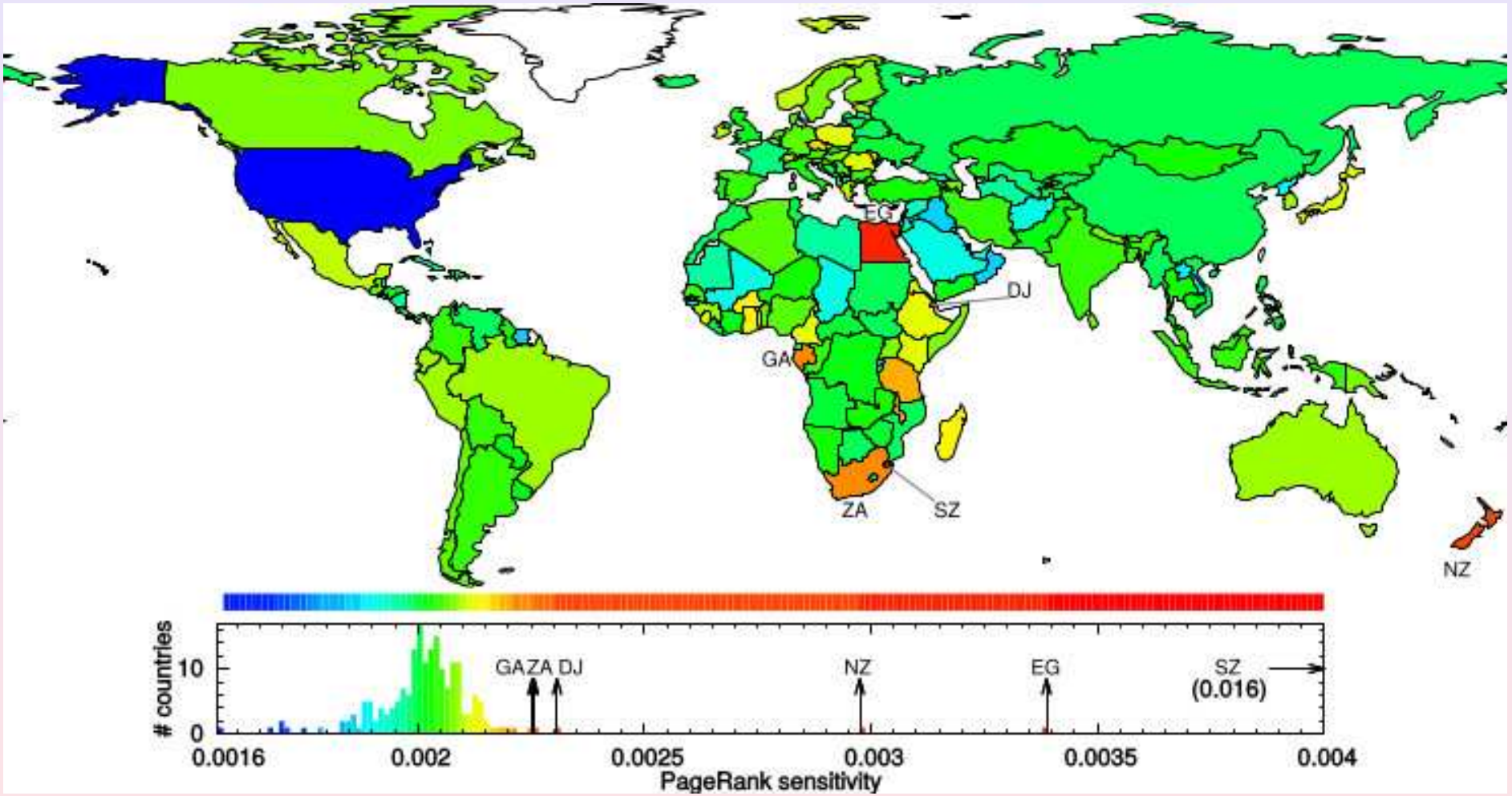
Network of close “friends”



Left: top category diseases and their 2 best “friend” (full circles);
Right: same but also 2 friend countries (open circles)

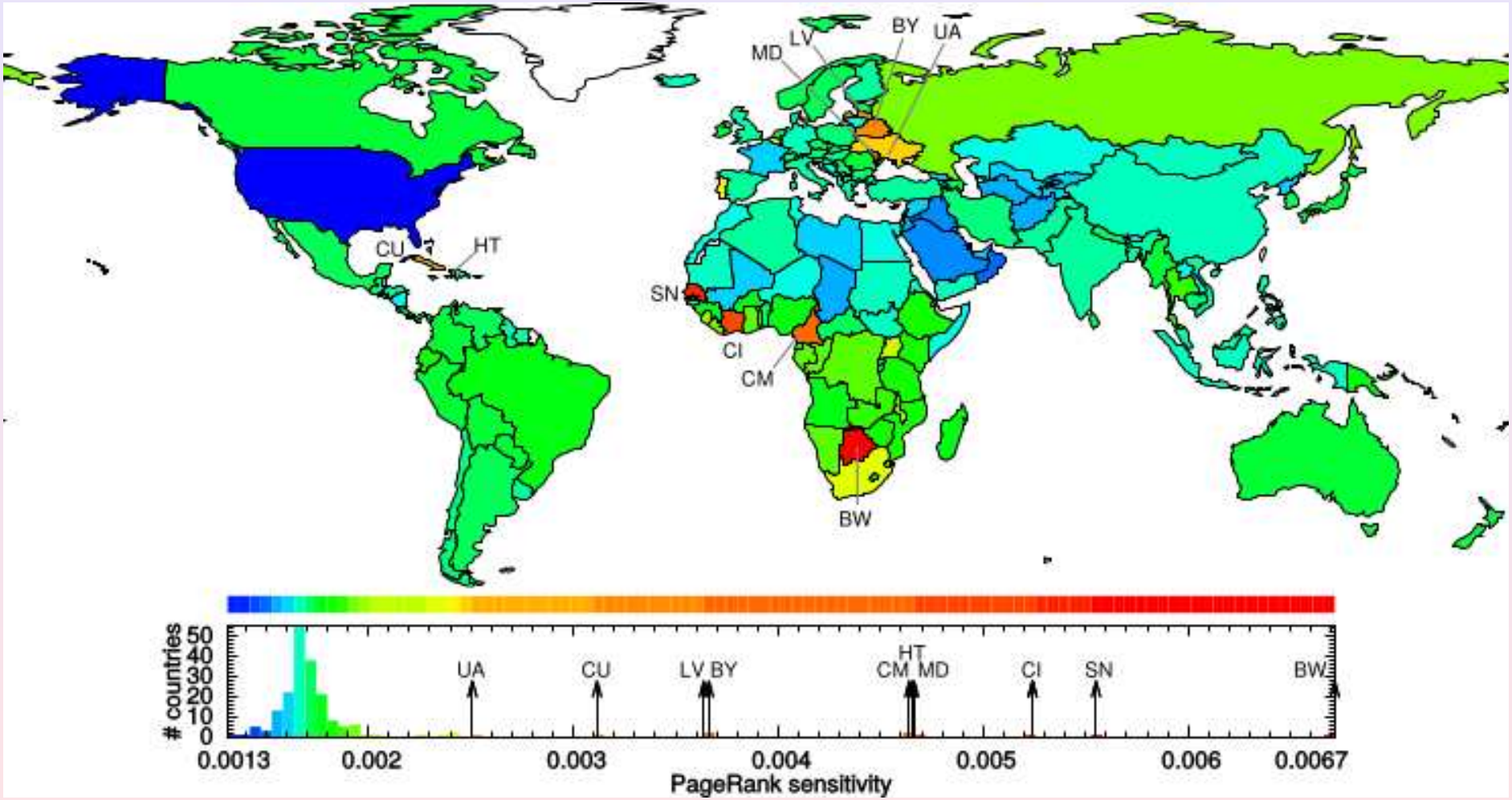
PageRank sensitivity of countries to tuberculosis

$S = d \ln(P_c)/d\delta_{tub}$ (weight variation disease-country)



PageRank sensitivity of countries to HIV/AIDS

$S = d\ln(P_c)/d\delta_{hiv}$ (weight variation disease-country)



Cancers, drugs and countries

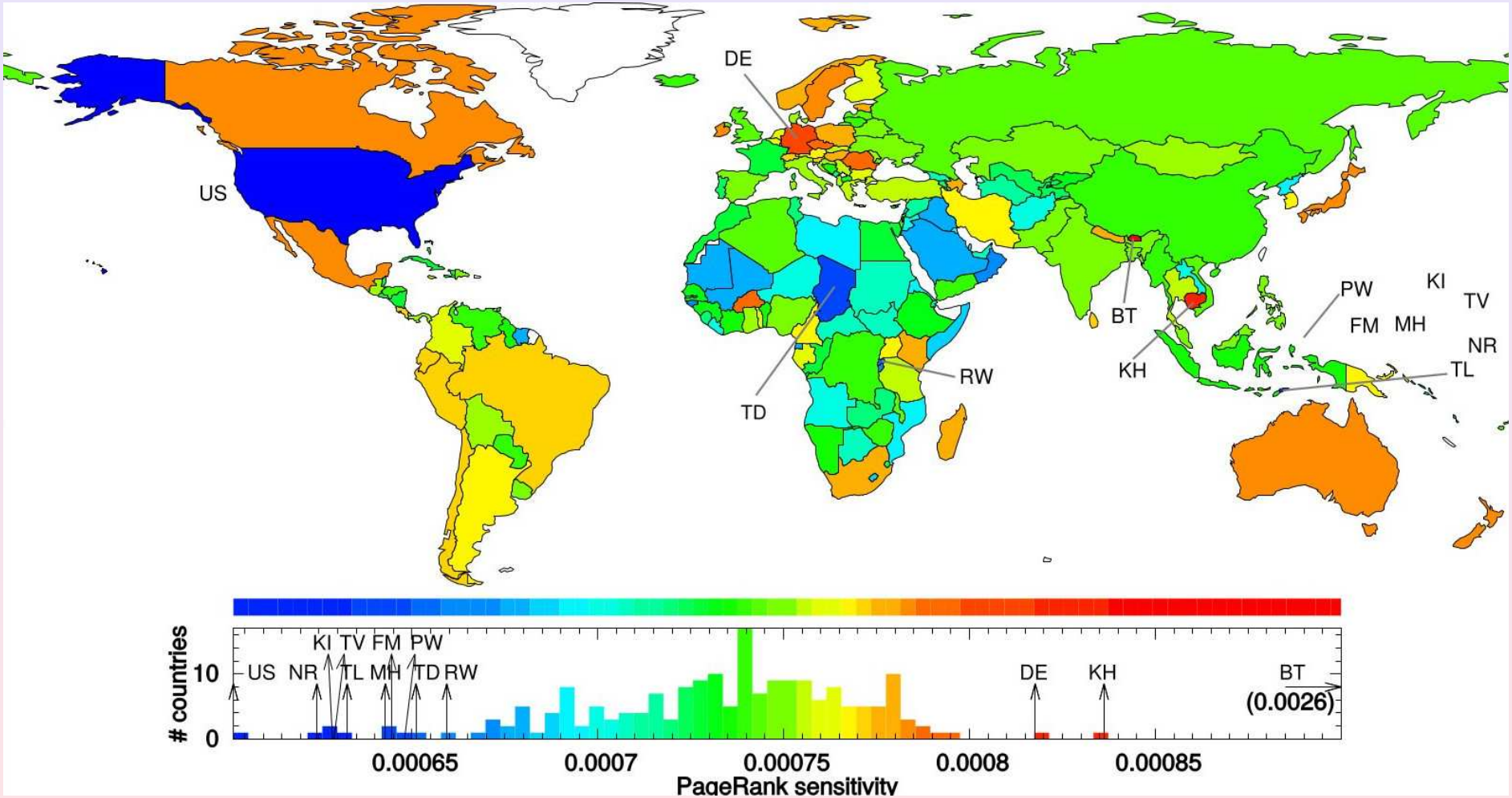
37 cancers, 203 drugs, 195 countries; PageRank order

Cancers and Drugs

Rank	Cancer	Drug
1	Lung cancer	Methotrexate
2	Breast cancer	Thalidomide
3	Leukemia	Paclitaxel
4	Prostate cancer	Prednisone
5	Colorectal cancer	Cisplatin
6	Brain tumor	Dexamethasone
7	Pancreatic cancer	Doxorubicin
8	Melanoma	Propranolol
9	Stomach cancer	Interleukin 2
10	Ovarian cancer	Cyclophosphamide

PageRank sensitivity of countries to lung cancer

$S = d\ln(P_c)/d\delta_{lung}$ (weight variation disease-country)



Further applications of Markov chains and Google matrix ?

→ Artificial intelligence for bionetworks

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