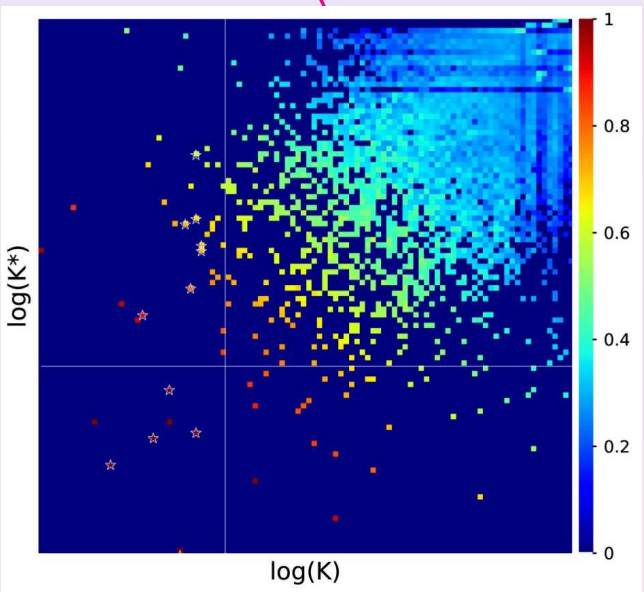
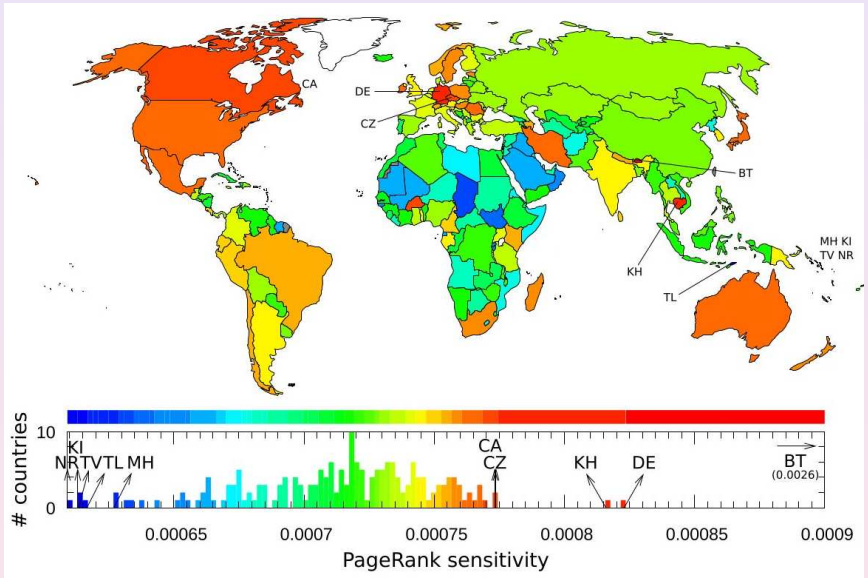
Google matrix analysis of protein-protein interactions from MetaCore and TRANSPATH networks

Dima Shepelyansky (CNRS-UPS, Toulouse)
www.quantware.ups-tlse.fr/dima
with E.Kotelnikova (Clarivate CAT), K.Frahm (UPS), J.Lages (U Besancon)

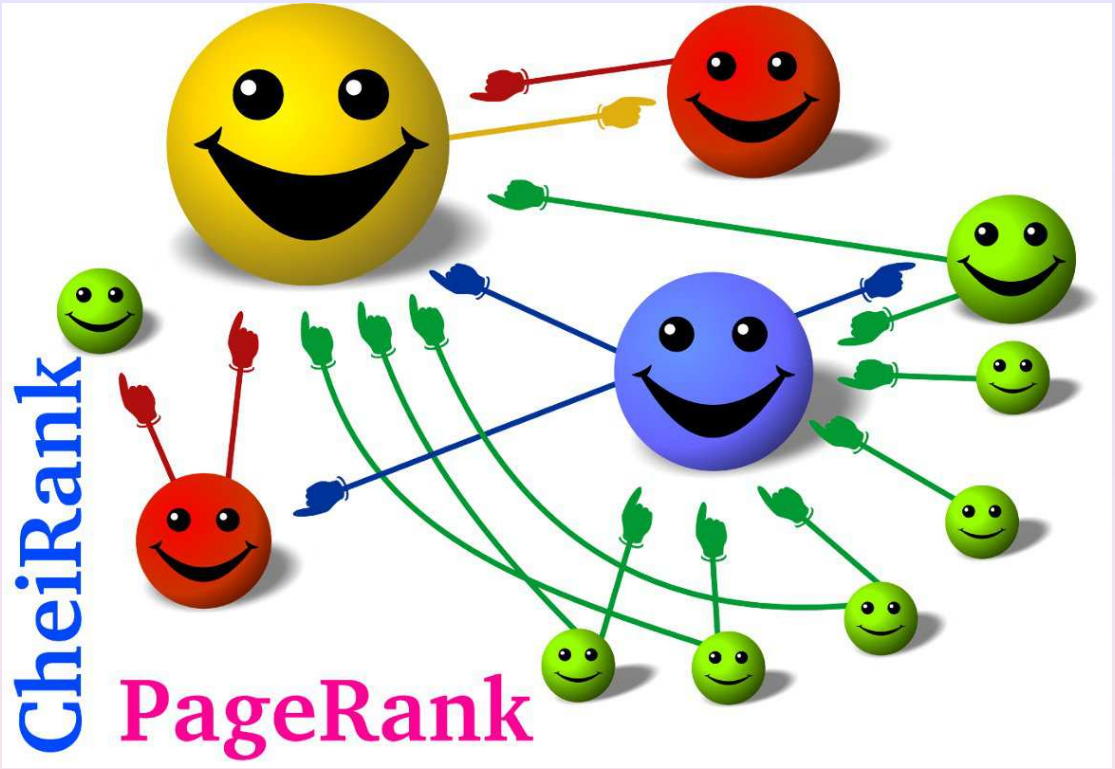


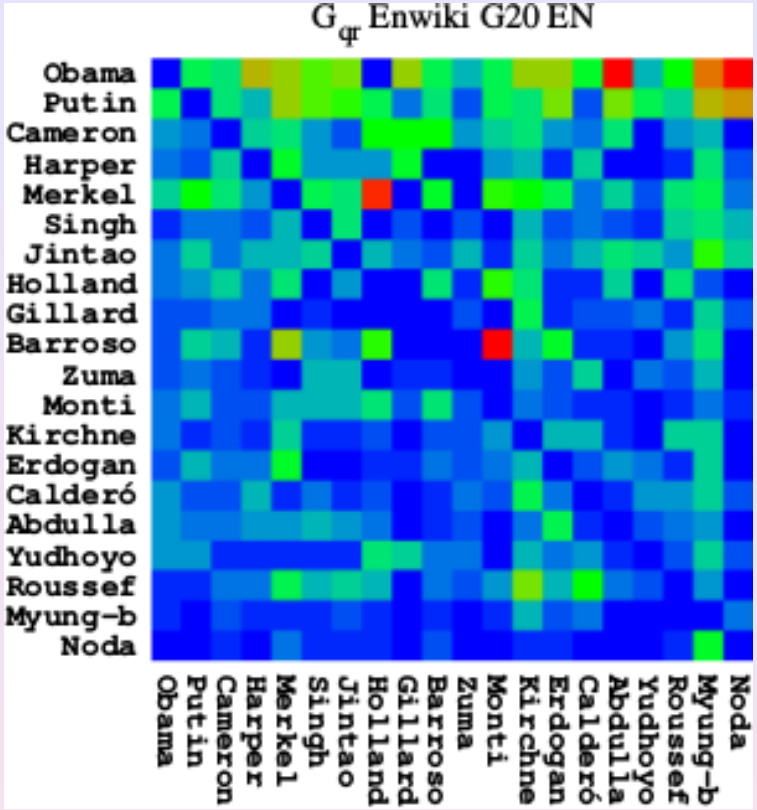


* Markov chains (1906) $\rightarrow$ Brin and Page (1998) $\rightarrow$ Google matrix and search engine
* reduced Google matrix of directed networks (brief introduction)
* Applications: multiproduct world trade network (UN COMTRADE), Wikipedia Ranking of World Universities (WRWU), protein-protein interactions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

Support: LABEX NANOX MTDINA; Ref: bioRxiv 4 April (2021)

(1906) Markov vs Wigner (1955)





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

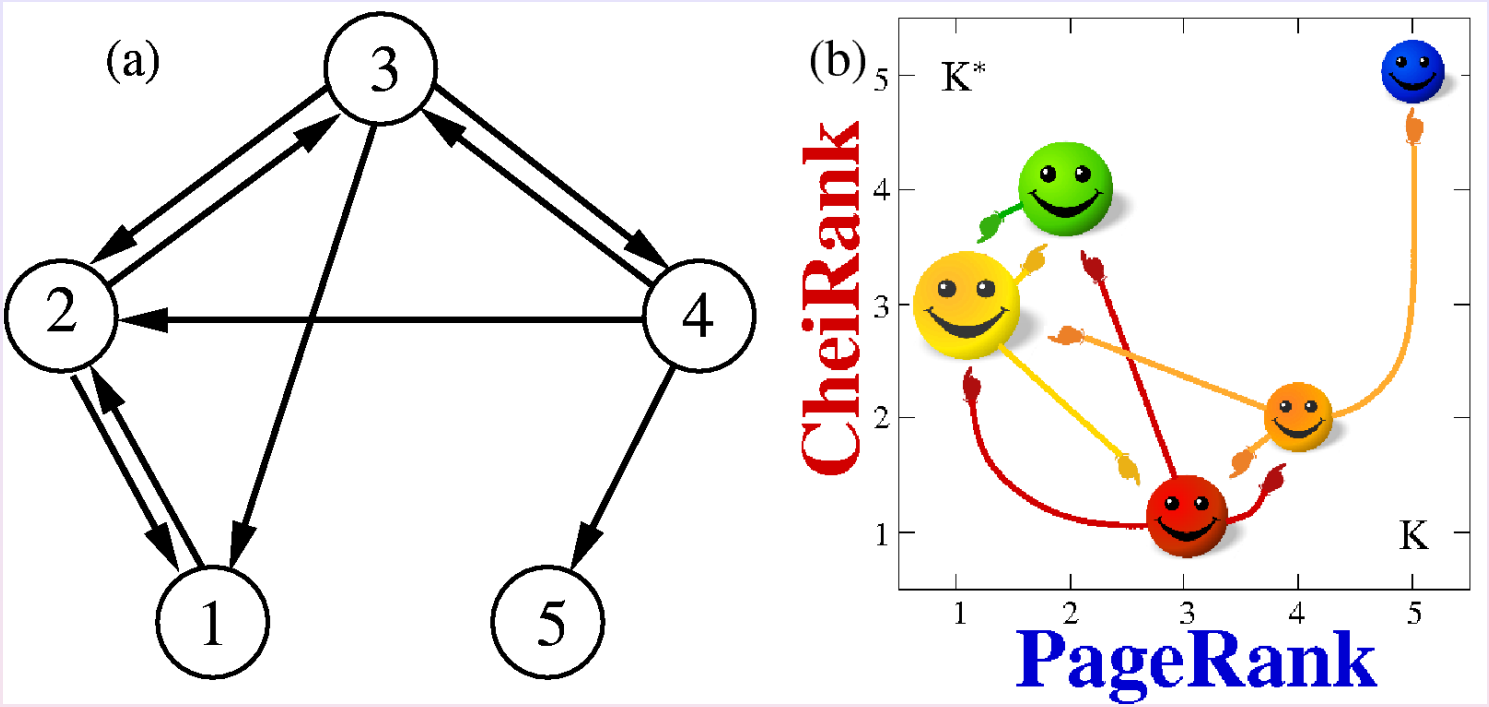
1991: WWW, small world social networks $\rightarrow$ Markov (1906) $\rightarrow$ Google matrix

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**

  $\mathbf{G} = \alpha \mathbf{S} + (1-\alpha)\frac{\mathbf{E}}{N}$; $\quad GP = \lambda P$ => 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) ...
  PageRank/CheiRank index $K/K^* \rightarrow$ monotonic decrease of probability $P(K)/P^*(K^*)$ [Ermann, Frahm, DS Rev. Mod. Phys. **87**, 1261 (2015)]

Computation algorithms

* PageRank and CheiRank vectors 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 (REGOMAX) $\rightarrow$ below

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: $N \approx 4000, 40000, 270000$

see Ermann, Frahm, DS Rev Mod Phys (2015)
http://www.quantware.ups-tlse.fr/dima/subjgoogle.html
Excellent mark for EC FET Open project NADINE (2012-2015) coordinated by Quantware (www.quantware.ups-tlse.fr/FETNADINE/)

Reduced Google matrix (REGOMAX)

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);
K.Frahm, K.Jaffrres-Runser, DS EPJB **89**, 269 (2016)

Protein-protein interactions for cancer networks

PLOS ONE | Reduced Google matrix for directed biological networks

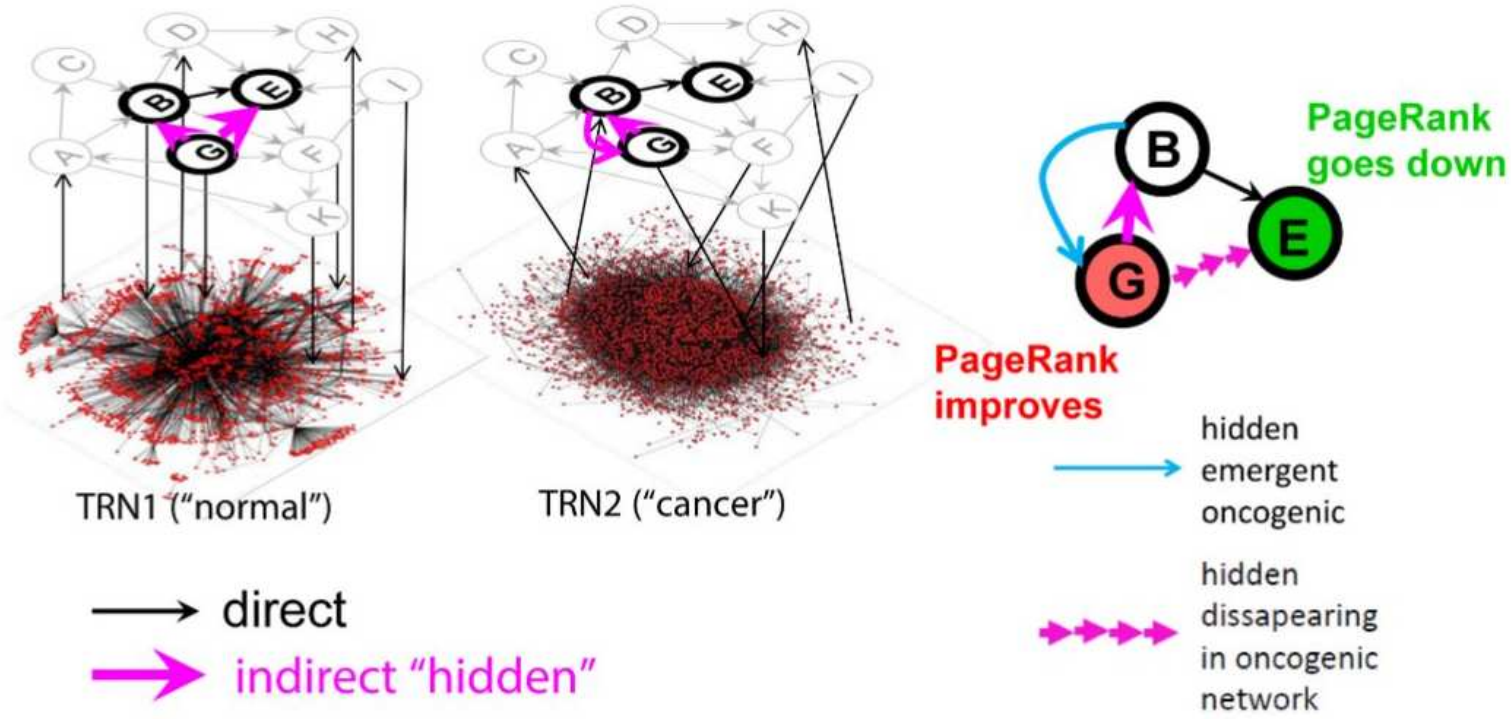


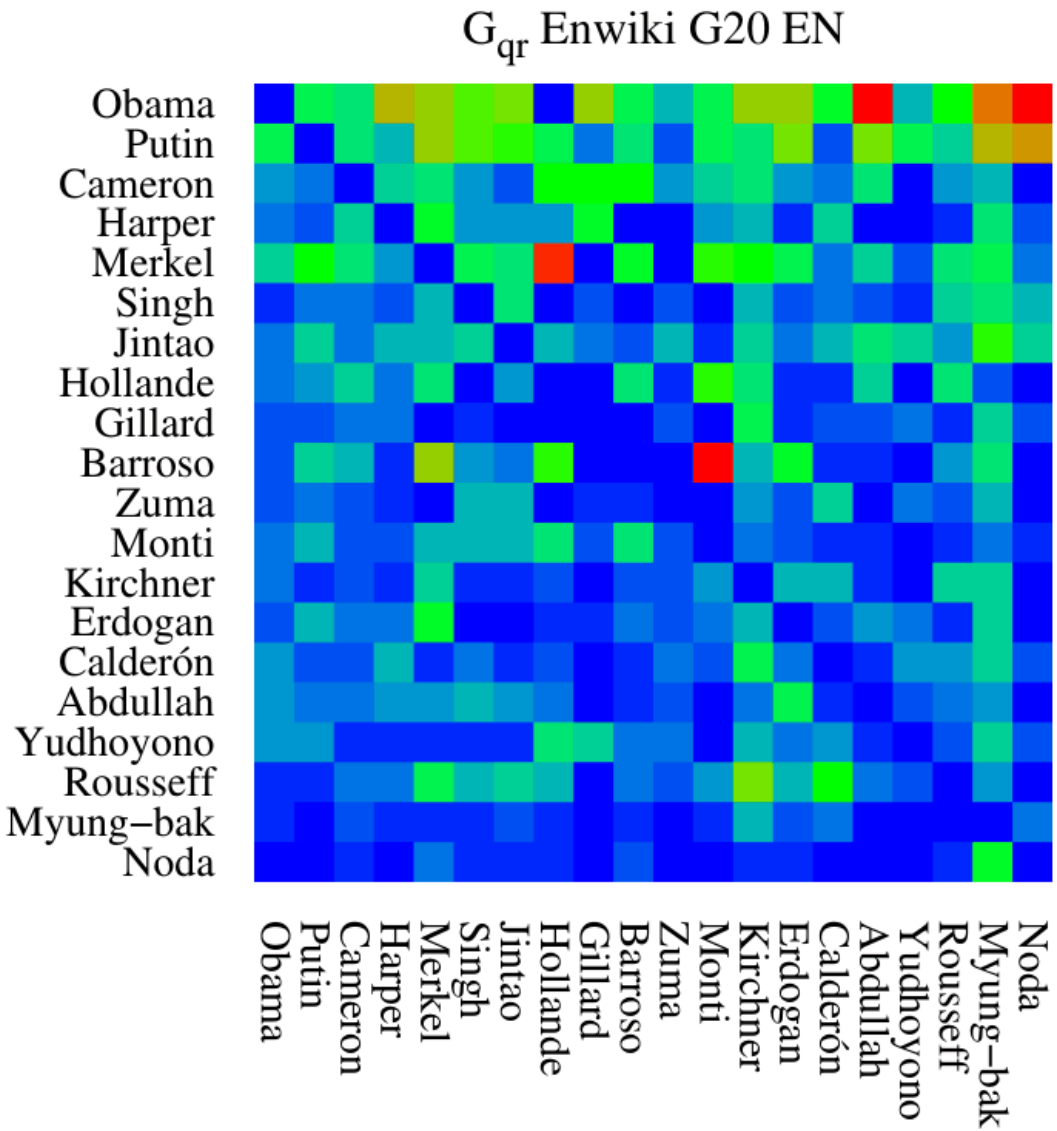
Fig 1. Using reduced Google matrix approach for inferring hidden causal relations in signaling pathways. Here the structure of the context-dependent global regulatory network is symbolically shown as consisting of two layers: the upper (nodes A-K) is the global signaling network whose structure does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

G-reduced: G20 political leaders 2012-ENWIKI2013

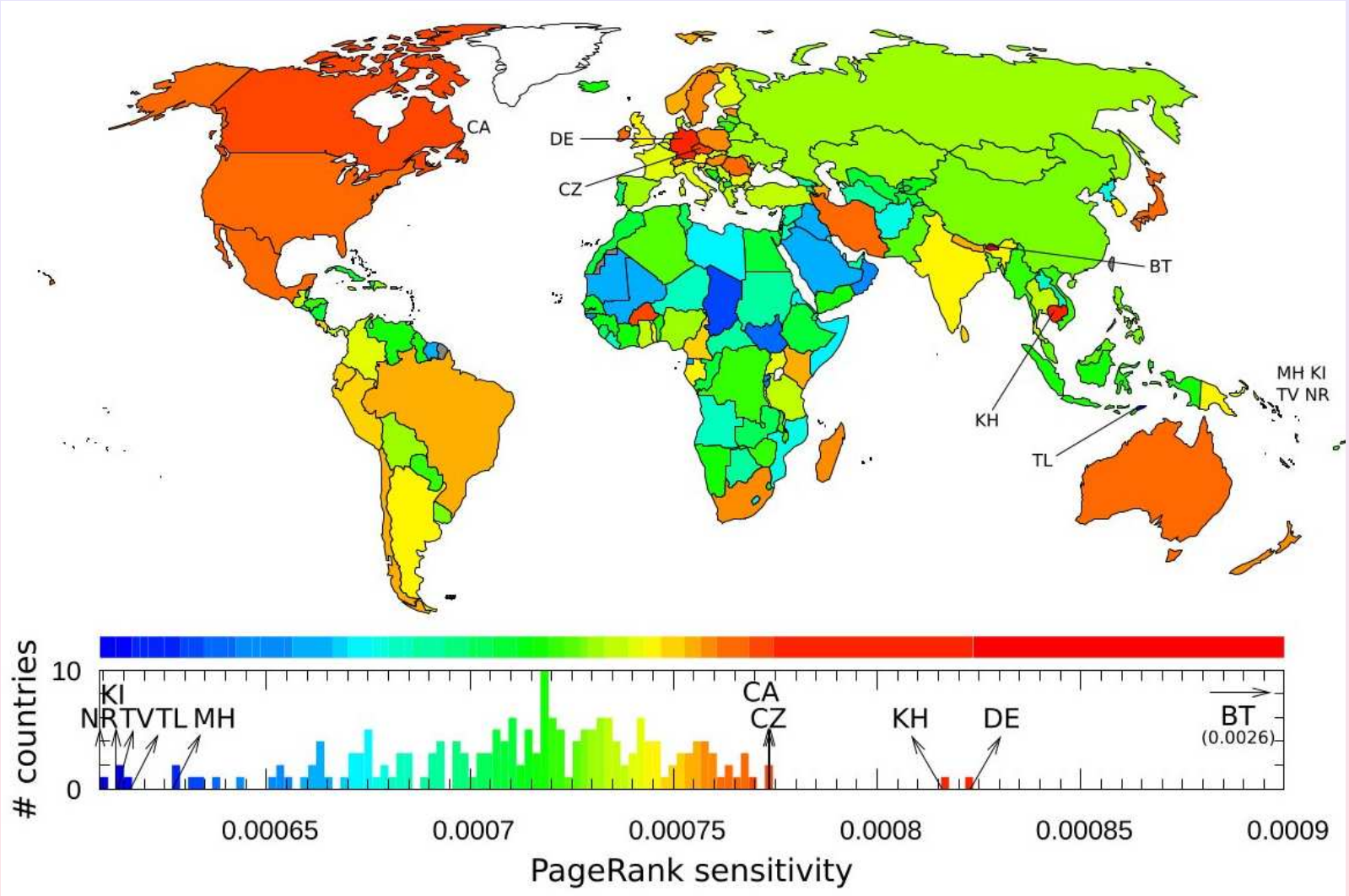
G_R example: G20 political leaders 2012 indirect links of G_{qr} (non-diagterms)



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

PageRank sensitivity of countries to lung cancer

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



195 countries (5.4/122 millions ENWIKI2017 articles/links)

Rollin, Lages, DS PLOS ONE **14(9)**, e0222508 (2019)

Databases of PPI networks and REGOMAX

* SIGNOR (Rome IT) public, $N = 4341$ proteins (nodes), $N_\ell = 12567$ links (April 2019); Frahm, DS Physica A **559**, 125019 (2020)

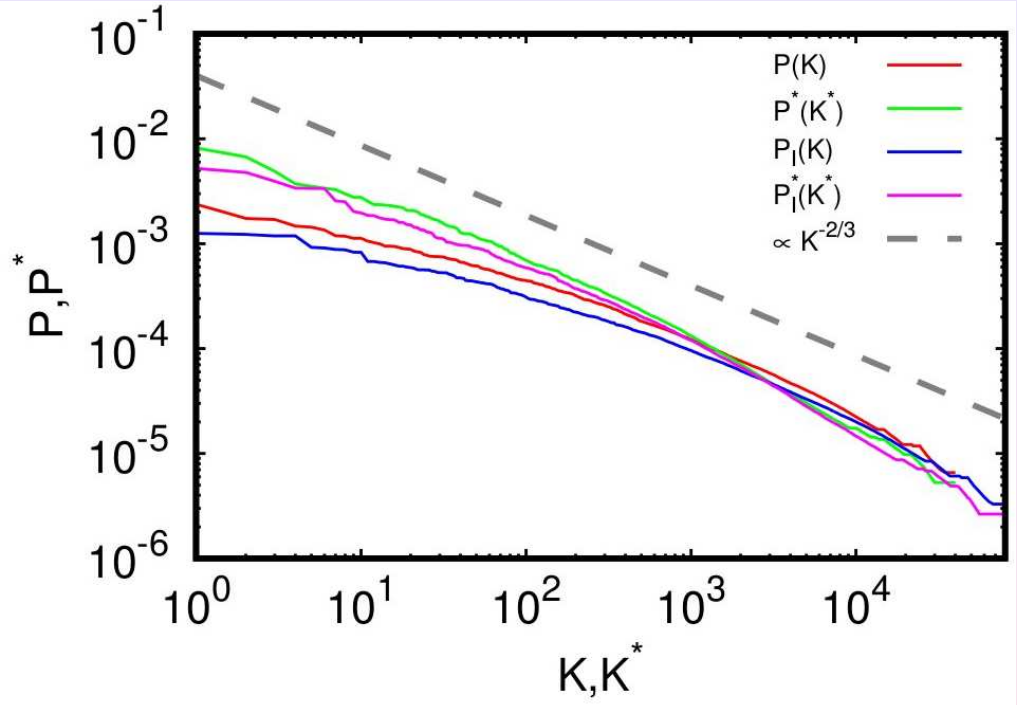


Figure 1 PageRank probability $P(K)$ ($P_I(K)$) and CheiRank probability $P^*(K^*)$ ($P^*_I(K)$) are shown as a function of the corresponding rank indexes K and K^* for the simple (Ising) MetaCore network. For comparison, the dashed gray line corresponds to the power decay $P \propto K^{-2/3}$.

* MetaCore (Clarivate CAT-UK) commercial (several Millions USD for global net), $N = 40079$ nodes, $N_\ell = 292904$ links; Kotelnikova, Frahm, Lages, DS bioRxiv April 4 (2021)

* TRANSPATH commercial (genexplain.com/transpath/), $N = 272455$ nodes, $N_\ell = 13797637$ links; Alexander Kel (DE), Frahm, DS in progress

Bi-functional nature of PPI networks: Ising spin

activation or inhibition action => doubling of node and Ising spin links

2.3 Bi-functional Ising MetaCore network

To take into account the bi-functional nature (activation and inhibition) of MetaCore links, we use the approach proposed in [9, 10] with the construction of a larger network where each node is split into two new nodes with labels (+) and (−). These two nodes can be viewed as two Ising-spin components associated to the activation and the inhibition of the corresponding protein. To construct the doubled "Ising" network of proteins, each elements of the initial adjacency matrix is replaced by one of the following 2×2 matrices

$$\sigma_+ = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \sigma_- = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \quad \sigma_0 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{6}$$

where σ_+ applies to "activation" links, σ_- to "inhibition" links, and σ_0 when the nature of the interaction is "unknown" or "neutral". For the rare cases of multiple

Frahm, DS Physica A (2020);
Kotelnikova, Frahm, Lages, DS bioRxiv April 4 (2021)

PageRank-CheiRank plane of MetaCore network

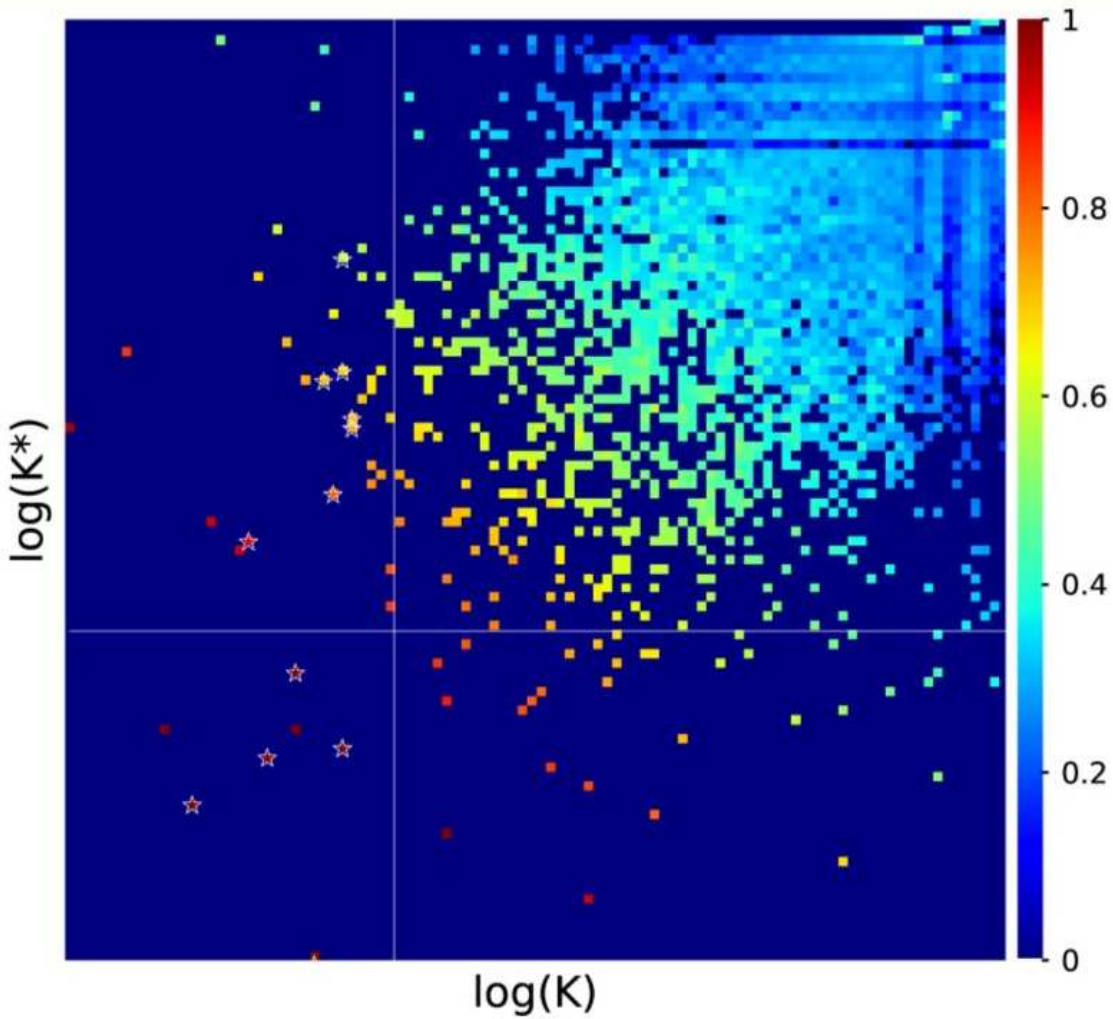


Figure 2 Density of nodes of the MetaCore network on the PageRank-CheiRank (K, K^*)-plane. The numbers of the color bar are a linear function of the logarithm of the density (with maximum values corresponding to 1 (red); minimum non-zero and zero values of the density corresponding to 0 (blue); the distribution is computed for 100×100 cells equidistant in logarithmic scale). The white stars indicate the positions of the 12 selected nodes presented in Table 1. The white vertical and horizontal lines represent nodes with $K \leq 40$ and $K^* \leq 40$, respectively.

“Magnetization” of proteins

$$M(K(j)) = (P_+(K(j)) - P_-(K(j)))/(P_+(K(j)) + P_-(K(j))); (-1 \leq M \leq 1)$$

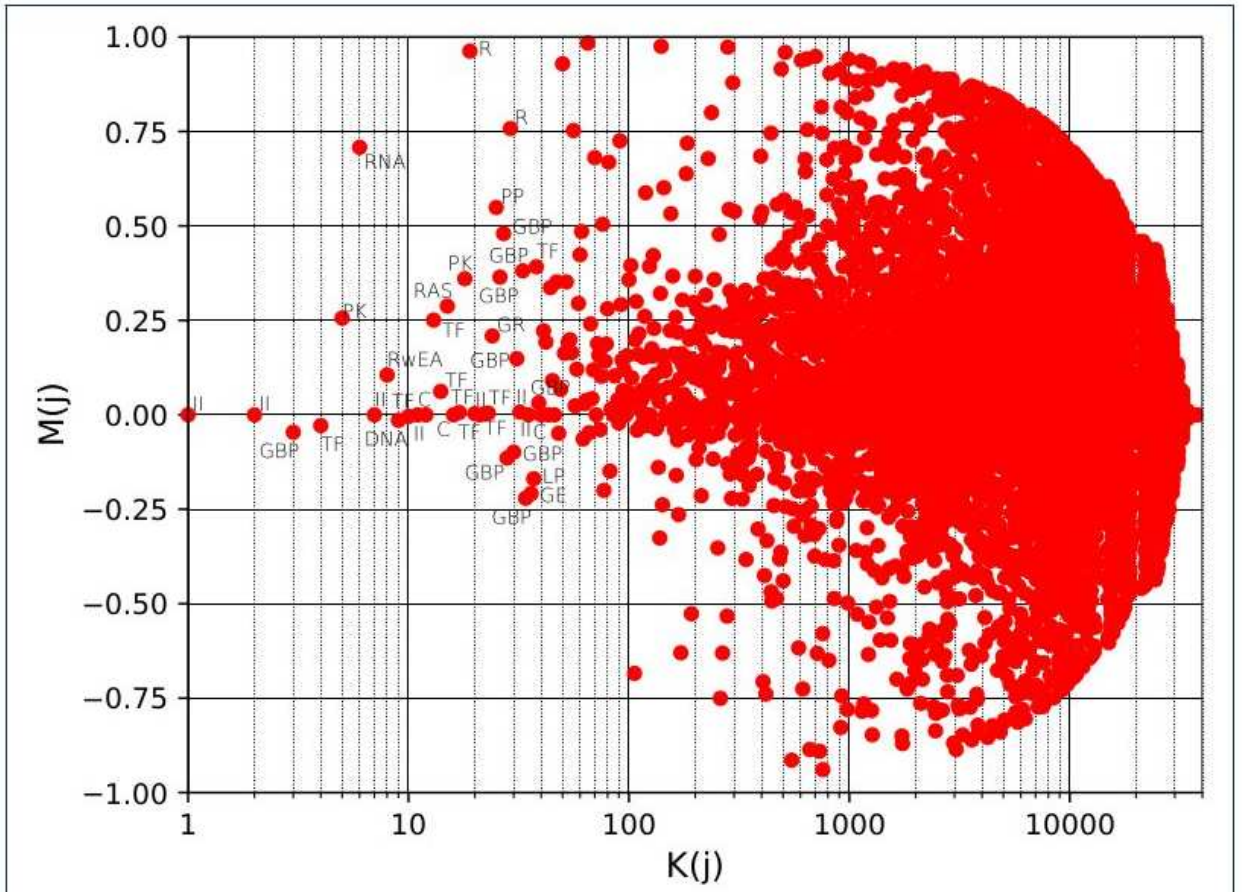


Figure 3 PageRank magnetization $M(j) = (P_+(j) - P_-(j))/(P_+(j) + P_-(j))$ for the Ising MetaCore network. Here, j is the node index and $K(j)$ is the PageRank index of the node j in the simple Metacore network (without node doubling). The biological class is reported for the top 40 PageRank nodes ($K \leq 40$, see Table 1): Inorganic ions (II), Generic binding protein (GBP), Transcription factor (TF), Protein kinase (PK), RNA, Receptor with enzyme activity (RwEA), DNA, Compound (C), RAS superfamily (RAS), Reaction (R), Generic receptor (GR), Protein phosphatase (PP), Generic enzyme (GE), Lipid phosphatase (LP).

Top 40 PageRank proteins of MetaCore

K	$P(K)$ (10^{-2})	k	$M(K)$	Name	Class	Localization
1	0.2506		0	H^+ cytosol	Inorganic ion	Cytosol
2	0.2376		0	Na^+ cytosol	Inorganic ion	Cytosol
3	0.1741		-0.045970	Beta-catenin	Generic binding protein	Cytoplasm
4	0.1701	1	-0.028308	p53	Transcription factor	Nucleus
5	0.1469		0.256018	c-Src	Protein kinase	Cytoplasm
6	0.1435		0.708154	mRNA intracellular	RNA	Intracellular
7	0.1352		0	H^+ extracellular region	Inorganic ion	Extracellular region
8	0.1189	2	0.105603	EGFR	Receptor with enzyme activity	Plasma membrane
9	0.1180		-0.014278	DNA	DNA	Nucleus
10	0.1125	3	-0.004135	ESR1 (nuclear)	Transcription factor	Nucleus
11	0.1125		0	K^+ extracellular region	Inorganic ion	Extracellular region
12	0.1056		0	ADP cytoplasm	Compound	Cytoplasm
13	0.1023	4	0.250910	STAT3	Transcription factor	Nucleus
14	0.0997		0.062046	Androgen receptor	Transcription factor	Nucleus
15	0.0947		0.287801	Rac1	RAS superfamily	Cytoplasm
16	0.0946		0	PO_4^{3-} cytoplasm	Compound	Cytoplasm
17	0.0940	5	0.006332	c-Myc	Transcription factor	Nucleus
18	0.0919	6	0.360271	FAK1	Protein kinase	Cytoplasm
19	0.0899		0.962815	cytosol $K^+ \rightarrow$ extracellular region K^+	Reaction	NA
20	0.0889	7	0.003377	ESR2 (nuclear)	Transcription factor	Nucleus
21	0.0884		0	K^+ cytosol	Inorganic ion	Cytosol
22	0.0849	8	0.002825	RelA (p65 NF-kB subunit)	Transcription factor	Nucleus
23	0.0834	9	0.004567	ARX	Transcription factor	Cytoplasm
24	0.0828	10	0.208984	ITGB1	Generic receptor	Plasma membrane
25	0.0787	11	0.548888	SHP-2	Protein phosphatase	Cytoplasm
26	0.0776	12	0.364614	GRB2	Generic binding protein	Cytoplasm
27	0.0760		0.479956	PI3K reg class IA (p85)	Generic binding protein	Cytoplasm
28	0.0759		-0.114311	E-cadherin	Generic binding protein	Plasma membrane
29	0.0754		0.757892	$CO_2 + H_2O \rightarrow H^+ + HCO_3^-$	Reaction	NA
30	0.0753		-0.098664	p21	Generic binding protein	Nucleus
31	0.0752		0.148707	Caveolin-1	Generic binding protein	Cytoplasm
32	0.0749		0.007470	Ca^{2+} cytosol	Inorganic ion	Cytosol
33	0.0744		0.381345	PI3K reg class IA (p85-alpha)	Generic binding protein	Cytoplasm
34	0.0727		-0.220751	Bcl-2	Generic binding protein	Mitochondrion
35	0.0720		0	Cl^- intracellular	Inorganic ion	Intracellular
36	0.0712		-0.208082	MDM2	Generic enzyme	Nucleus
37	0.0707		-0.169004	PTEN	Lipid phosphatase	Cytoplasm
38	0.0702		0.391984	PPAR-gamma	Transcription factor	Nucleus
39	0.0698		0.031543	ACTB	Generic binding protein	Cytoplasm
40	0.0679		0	Acetyl-CoA intracellular	Compound	Intracellular

Top 40 CheiRank proteins of MetaCore

K^*	$P^*(K^*)$ (10^{-2})	k^*	$M(K^*)$	Name	Class	Localization
1	1.1464	1	0.006332	c-Myc	Transcription factor	Nucleus
2	0.8172		0.035667	eIF2C2 (Argonaute-2)	Generic enzyme	Cytoplasm
3	0.6722		-0.174071	IGF2BP3	Generic binding protein	Cytoplasm
4	0.4890		0.680968	Ubiquitin	Generic binding protein	Cytoplasm
5	0.3719		0.110759	SOX9	Transcription factor	Nucleus
6	0.3529	2	-0.028308	p53	Transcription factor	Nucleus
7	0.3373		0.228978	c-Fos	Transcription factor	Nucleus
8	0.3276		0	CUX1 (p110)	Transcription factor	Nucleus
9	0.2989		-0.057557	SP1	Transcription factor	Nucleus
10	0.2770	3	-0.004135	ESR1 (nuclear)	Transcription factor	Nucleus
11	0.2769	4	0.002825	RelA (p65 NF-kB subunit)	Transcription factor	Nucleus
12	0.2534		-0.010911	eIF2C1 (Argonaute-1)	Generic binding protein	Cytoplasm
13	0.2354		0.062046	Androgen receptor	Transcription factor	Nucleus
14	0.2350		-0.045970	Beta-catenin	Generic binding protein	Cytoplasm
15	0.2330		-0.075622	BRD4	Generic binding protein	Nucleus
16	0.2308		0.153950	Oct-3/4	Transcription factor	Nucleus
17	0.2259		-0.001577	PUM2	Generic binding protein	Cytoplasm
18	0.2239		0.188479	EZH2	Generic enzyme	Nucleus
19	0.2193		0.208146	p300	Generic enzyme	Nucleus
20	0.2072		-0.407833	TUG1	RNA	Cytoplasm
21	0.2072		-0.118501	E2F1	Transcription factor	Nucleus
22	0.2062		0	ASCC2	Generic binding protein	Nucleus
23	0.2005		0	LIMR	Generic receptor	Plasma membrane
24	0.1903		0.148471	BRG1	Generic enzyme	Nucleus
25	0.1871	5	0.250910	STAT3	Transcription factor	Nucleus
26	0.1811		0.381258	RBM24	Generic binding protein	Cytoplasm
27	0.1789		0.746981	SUMO-1	Generic binding protein	Nucleus
28	0.1728		0.140357	c-IAP2	Generic binding protein	Cytoplasm
29	0.1699		0.038221	HIF1A	Transcription factor	Nucleus
30	0.1677		0	Zn^{2+} cytosol	Inorganic ion	Cytosol
31	0.1623		-0.013644	CDK9	Protein kinase	Cytoplasm
32	0.1587		-0.223816	MeCP2	Generic binding protein	Nucleus
33	0.1533		-0.053592	ELAVL1 (HuR)	Generic binding protein	Nucleus
34	0.1497		0.120649	HDAC1	Generic enzyme	Nucleus
35	0.1473		-0.034082	BRD7	Generic binding protein	Nucleus
36	0.1452		0.131956	CREB1	Transcription factor	Nucleus
37	0.1449		0	Zn^{2+} nucleus	Inorganic ion	Nucleus
38	0.1423		0.096830	SUMO-2	Generic binding protein	Cytoplasm
39	0.1400		-0.051730	BRD2	Protein kinase	Cytoplasm
40	0.1343		0.228824	C/EBPbeta	Transcription factor	Nucleus

Reduced Google matrix of 12 proteins

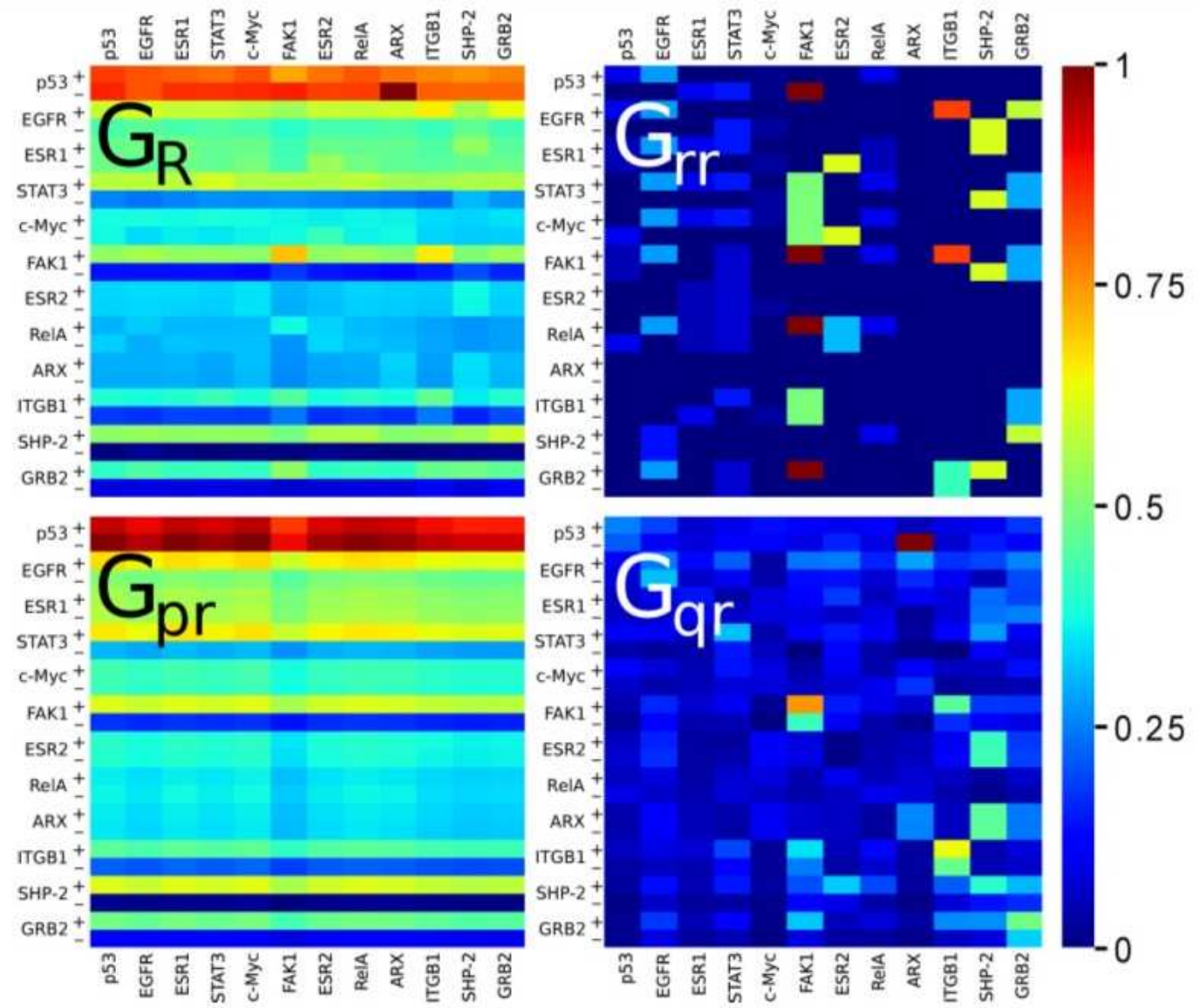


Figure 4 Reduced Google matrix G_R and its three matrix components G_{pr}, G_{rr} and G_{qr} associated to the subset of nodes presented in Table 1 and belonging to the Ising MetaCore network. The weights of the matrix components are $W_{pr} = 0.952$, $W_{rr} = 0.015$, and $W_{qr} = 0.033$.

2 components of REGOMAX of 12 proteins

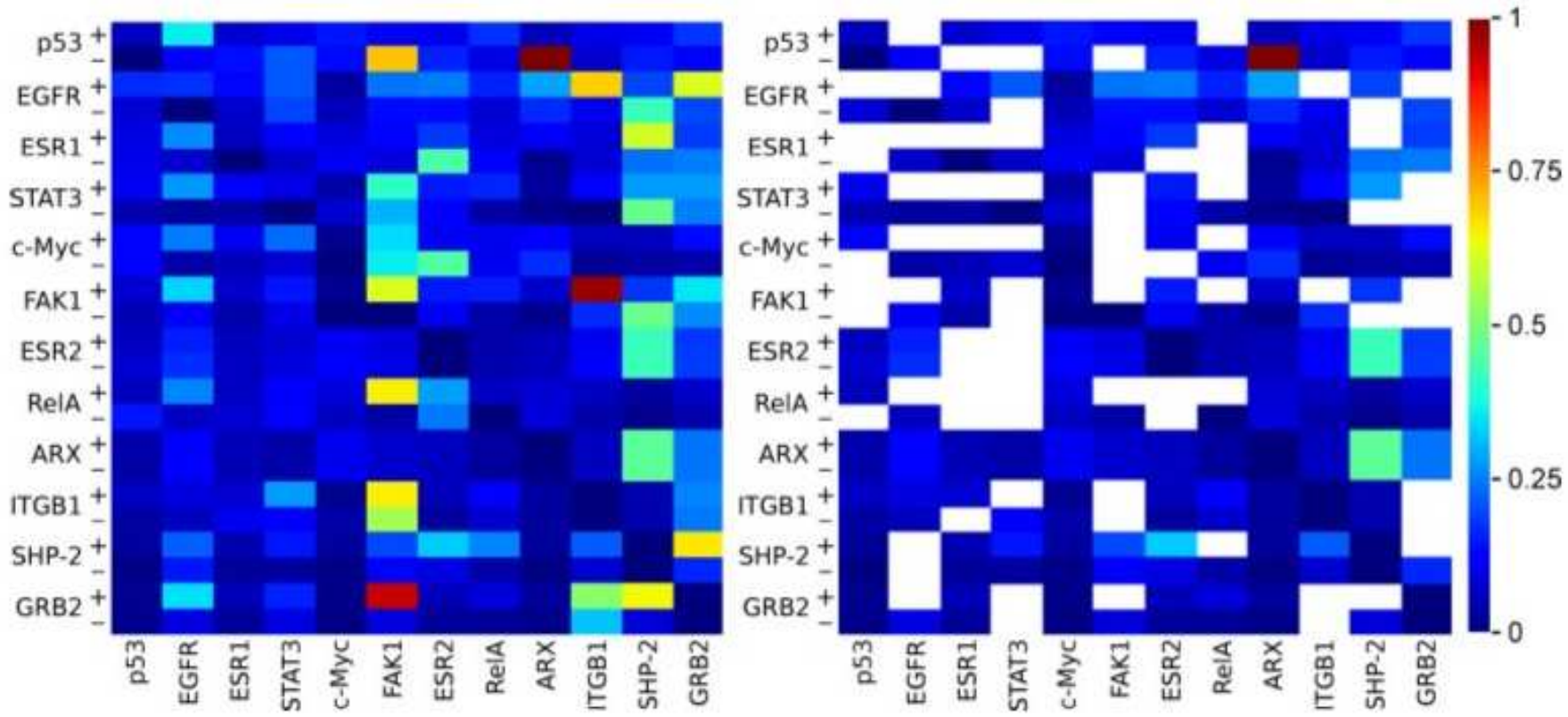


Figure 5 Sum of the two matrix components $G_{rr} + G_{qr}^{(\text{nd-block})}$. The matrix components are the same as in Fig. 4 with the exception of $G_{qr}^{(\text{nd-block})}$ which is obtained from G_{qr} by excluding 2×2 diagonal blocks, each one of these blocks corresponding to a protein self-loop. The right panel is the same as the left panel with the exception of the white cells which hide the direct links $j \to i$ between the 12×2 chosen nodes in the Ising MetaCore network. The values of the color bar correspond to the ratio of the matrix element over its maximum value.

PPI network structure

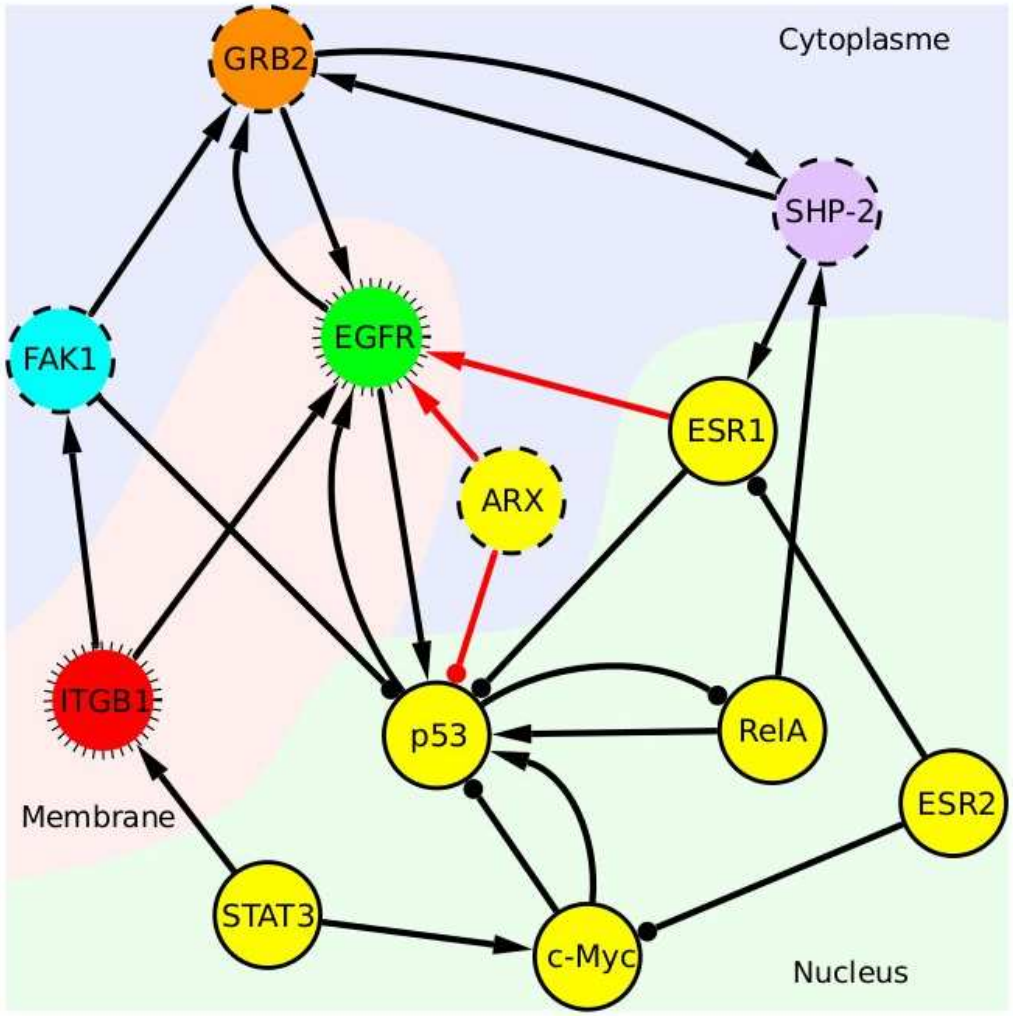


Figure 6 Reduced network of the chosen twelve proteins (see Table 1). The construction procedure of this network is given in the main text. Arrow edges (→) represent activation links and dot edges (⊸) represent inhibition links. The arrow tips and the dots are on the side of the target nodes. The black edges represent direct links. The red edges represent hidden links. The color of the nodes correspond to the type of proteins: transcription factors (yellow), protein kinase (cyan), generic receptor (red), receptor with enzyme activity (green), general binding protein (orange), and protein phosphatase (violet). The border style of the node correspond to the location of the proteins: nucleus (solid line), cytoplasm (dashed line), and plasma membrane (hairy line).

PageRank sensitivity of proteins

$D_{(b\to a}(j) = d \ln P_{r\varepsilon}(j)/d\varepsilon$; $D_{(a\leftrightarrow b)}(j) = D_{(b\to a)}(j) + D_{(a\to b)}(j)$
$F_+(a,b) = D_{(a\leftrightarrow b)}(a) + D_{(a\leftrightarrow b)}(b)$; $F_-(a,b) = D_{(a\leftrightarrow b)}(a) - D_{(a\leftrightarrow b)}(b)$

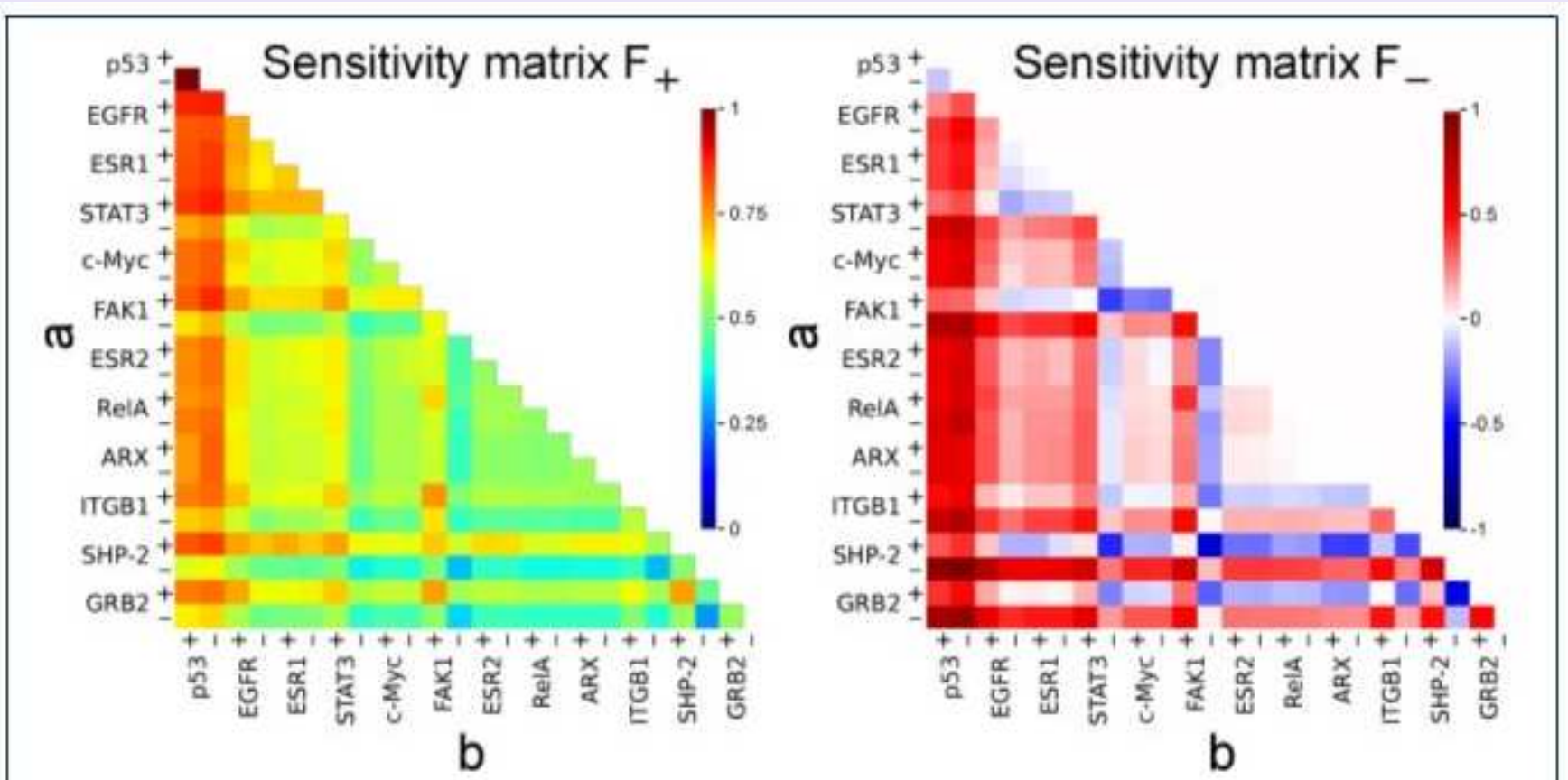


Figure 7 PageRank sensitivity matrices $F_+(a,b)$ (left panel) and $F_-(a,b)$ (right panel) associated to the subset of nodes presented in Table 1 and belonging to the Ising MetaCore network. The values of the color bar correspond to $F_+(a,b)/\max_{a,b}(F_+(a,b))$ (left panel) and to $F_-(a,b)/\max_{a,b}|F_-(a,b)|$ (right panel).

Magnetization of 12 proteins

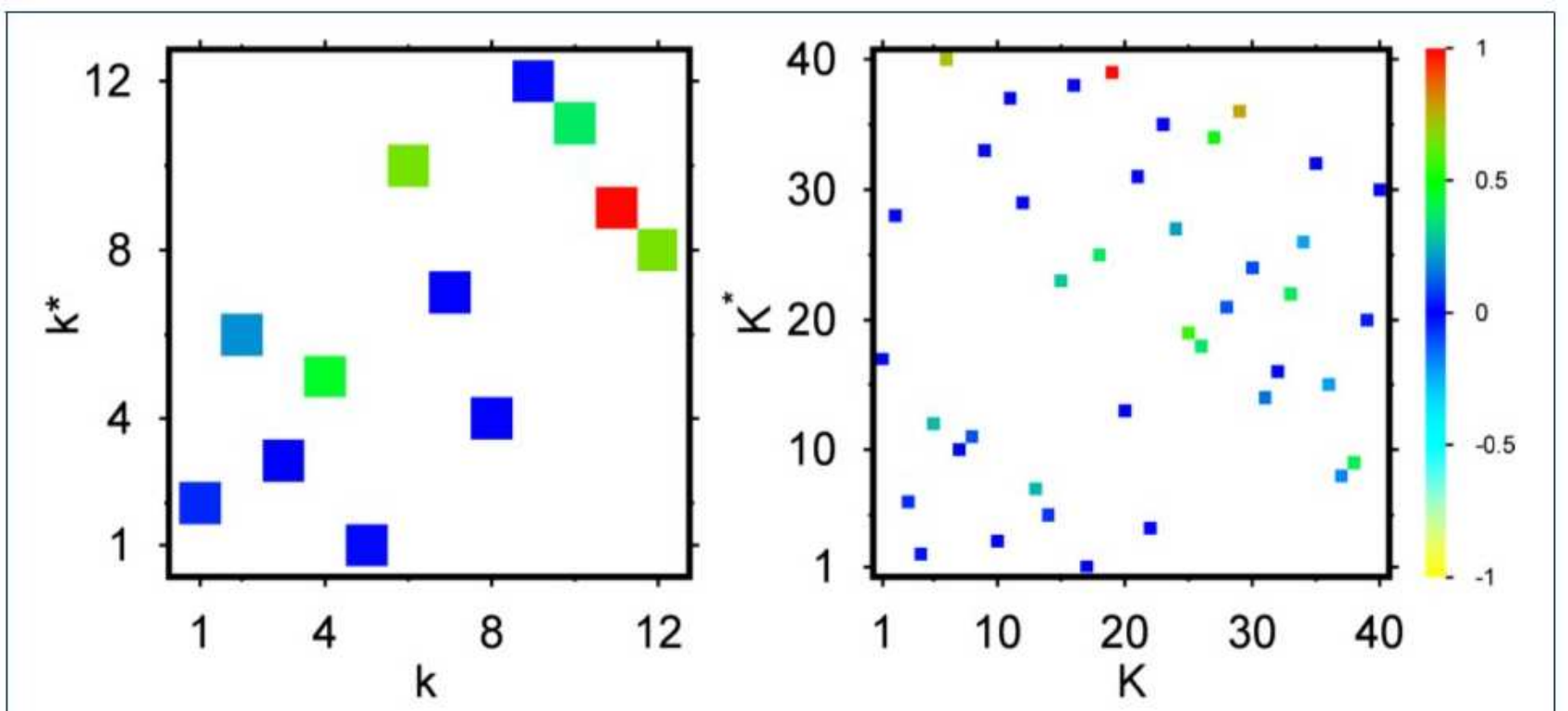


Figure 8 PageRank magnetization $M(K) = (P_+(K) - P_-(K))/(P_+(K) + P_-(K))$ presented in the PageRank-CheiRank (K, K^*)-plane. Left panel: PageRank magnetization $M(k)$ for the chosen twelve proteins presented in the relative indexes (k, k^*)-plane (see k adn k^* indexes in Table 1). Right panel: PageRank magnetization $M(K)$ for nodes with $K \leq 40$. Here, $P_\pm(K)$ is the PageRank probability of the $(\pm)$ component of the Ising MetaCore network node associated with the K PageRank (see text). The values of the color bar correspond to $M/\max|M|$ with $\max_{k\leq 12}|M(k)| = 0.549$ (left panel) and $\max_{K\leq 40}|M(K)| = 0.963$ (right panel). On the right panel, the K^* index is here the relative CheiRank index inside the set of the first $K \leq 40$ nodes.

Discussion

Interdisciplinary REGOMAX4PPI project:

Medicine group (J.Mazieres CRCT)

Bio-PPI-network group (V.Pancaldi CRCT)

MetaCore group (E.Kotelnikova Clarivate Barcelona CAT)

TRANSPATH group (A.Kel genexplain.com DE)

REGOMAX group (DS Quantware LPT)