Interactions of pharmaceutical companies with world countries, cancers and rare diseases from Wikipedia network analysis

Guillaume Rollin[1], José Lages[1*], Tatiana S. Serebriyskaya[2], Dima L. Shepelyansky[3]

[1] Institut UTINAM, CNRS, UMR 6213, OSU THETA, Université de Bourgogne Franche-Comté, Besançon, France
[2] Laboratory for Translational Research and Personalized Medicine, Moscow Institute of Physics and Technology, Moscow, Russia
[3] Laboratoire de Physique Théorique, IRSAMC, Université de Toulouse, CNRS, UPS, 31062 Toulouse, France

* Corresponding author
jose.lages@utinam.cnrs.fr

Abstract

Using the English Wikipedia network of more than 5 million articles we analyze interactions and interlinks between the 34 largest pharmaceutical companies, 195 world countries, 47 rare renal diseases and 37 types of cancer. The recently developed algorithm using a reduced Google matrix (REGOMAX) allows us to take account both of direct Markov transitions between these articles and also of indirect transitions generated by the pathways between them via the global Wikipedia network. This approach therefore provides a compact description of interactions between these articles that allows us to determine the friendship networks between them, as well as the PageRank sensitivity of countries to pharmaceutical companies and rare renal diseases. We also show that the top pharmaceutical companies in terms of their Wikipedia PageRank are not those with the highest market capitalization.

Introduction

The improvement of human health and the treatment of various diseases is a vital task for human society [1]. The creation of efficient medications and drugs is now mainly controlled by large biotechnology and pharmaceutical companies. The world's 34 largest companies are listed in Wikipedia [2]. The analysis of interactions between these companies and their influence on various diseases is an important, but not an easy task. Here, we develop a data mining approach to this task by using the directed network of articles in the English Wikipedia, dated at May 2017, that is generated by citation links between articles. At present, Wikipedia is a repository of a huge amount of human knowledge, exceeding that of the Encyclopedia Britannica in terms of the volume and accuracy of articles devoted to scientific topics [3]. Scientific articles are actively maintained, as it can be seen by the example of articles on biomolecules [4]. Academic research into and analysis of the information present in Wikipedia are growing as new tools and methods are developed, as reviewed in [5]. The quality of Wikipedia articles is improving with time as shown by the analysis reported in [6].

At present a variety of methods exist for the analysis and characterization of complex networks (see e.g. [7]). The most popular analytical tool is the PageRank

algorithm invented by Brin and Page in 1998 for ranking the sites on the World Wide Web (WWW) [8]. A detailed description of the algorithm and the mathematical properties of the related Google matrix are given in [9] while a variety of applications of the Google matrix to real directed networks are described in [10]. The application of Google matrix methods to Wikipedia networks in 24 different languages has led to a reliable ranking of historical figures over 15 centuries of human history [11], as well as to the ranking of the world's universities [12].

Recently, the reduced Google matrix (REGOMAX) algorithm has been proposed for the analysis of the effective interactions and links between a selected subset of nodes of interest embedded in a much larger global network [13]. The REGOMAX algorithm originates from the scattering theory of nuclear and mesoscopic physics and the field of quantum chaos. Its efficiency has been demonstrated for Wikipedia networks retrieving known effective interactions between politicians [14], geopolitical relations and links between world countries [15], universities [16] and banks [17].

The REGOMAX approach has also been used for networks of protein-protein interactions [18] and for Wikipedia networks in order to determine the influence in Wikipedia of infectious diseases [19], drugs and cancers [20]. Here, we extend this approach to analyze the influence and mutual interactions in Wikipedia of the world's 34 largest biotechnology and pharmaceutical companies [2]. As the PageRank and the REGOMAX algorithms provide rankings of items according to how these items are connected, an item will, in the following, be designated as influential in Wikipedia if its ranking, mirroring the cascade of directed links converging toward it, is high. We use the Wikipedia network with 5 416 537 articles and establish relations between 195 countries and these companies on the basis of their PageRank sensitivity (see section Methods) of the corresponding articles. We also construct interaction networks between these companies and 47 rare renal diseases and 37 types of cancer. The number of people suffering from cancer in 2012 was about 32.6 million worldwide, with this number increasing by 6-9 million people each year [21]. By 2019, the number of people worldwide suffering from rare diseases had reached 350 million [22]. Despite the fact that cancer and 80% of rare diseases are genetic diseases, the challenge involved in developing drugs for these two groups of diseases differ. One of the fundamental challenge in the field of drug development for rare diseases is the huge number and diversity of such diseases. There are more than 7 000 known rare diseases [22] whereas the number of known cancer types is quite limited. The large number and diversity of rare diseases cause problems at the level of drug development due to the small number of patients accessible for individual disease, the logistics involved in reaching widely dispersed patient populations, the lack of validated biomarkers and surrogate end-points, and the limited clinical expertise and centers of excellence available [23]. The representation of information about cancer and rare diseases in Wikipedia clearly illustrates the difference between cancers and rare diseases. Most of the cancer types listed on the National Cancer Institute site [24] are represented in Wikipedia. At the same time, we observe that only 8.5% and 15% of the phenotypes described in Orphanet [25] and OMIM [26], respectively, are addressed by topics in Wikipedia. For our calculations, we use only the 47 rare renal diseases listed in [27] and represented in Wikipedia.

We hope that, through the prism of Wikipedia, our results will make it possible to gain a better understanding of the influence of pharmaceutical companies on these diseases and to obtain specialization profiles for these pharmaceutical companies.

The paper is constructed as follows: the mathematical methods are described in the Section Methods and data are overviewed in the Section Datasets. The two following Sections are devoted to the presentation of Results and of the Discussion.

Methods

Google matrix construction

A detailed description of the construction of the Google matrix and its properties can be found in [9, 10]. Here, we therefore provide only a brief description in order to assist readers.

Let us consider a network of N nodes described by the adjacency matrix A the elements of which are either $A_{ij} = 1$ if node j points to node i, or $A_{ij} = 0$ otherwise. The elements of the Google matrix G have the standard form $G_{ij} = \alpha S_{ij} + (1-\alpha)/N$ [8–10], where the matrix S describes Markov transitions from node to node. For a random surfer, the matrix element S_{ij} gives the probability of transition from node j to node i according to the directed network structure. The elements of matrix S are either $S_{ij} = A_{ij}/k_{out}(j)$ if the out-degree of node j, $k_{out}(j) = \sum_{i=1}^{N} A_{ij}$, is non zero, or $S_{ij} = 1/N$ if j has no outgoing links (dangling node), i.e., $k_{out}(j) = 0$. Hence, the random surfer can jump to any node once it reaches a dangling node. Here α is the damping factor ($0 < \alpha < 1$) and we use its standard value $\alpha = 0.85$ [9] in the following. It should be noted that for the range $0.5 \leq \alpha \leq 0.95$, the results are not sensitive to α [9, 10]. At each jump, the random surfer either follows the network structure with a transition probability proportional to α, or randomly chooses any node with a transition probability proportional to $1-\alpha$. The damping factor α prevents the random surfer from being trapped in dangling subnetworks [9].

Starting its journey from node j_0, after n iterations the random surfer will arrive at node i with the probability $p_i^{(n)}$, which is given by the ith component of the probability vector $p^{(n)} = G^n p^{(0)}$ where $\{p_i^{(0)} = 0\}_{i \neq j_0}$ and $p_{j_0}^{(0)} = 1$. According to the Perron-Frobenius theorem [9], for any initial vector $p^{(0)}$, the iteration process converges to a unique vector P, which, within the framework of the Google matrix is called the PageRank vector, and is given by the following limit $P = \lim_{n\to\infty} G^n v$ for any vector v with positive elements and $\sum_{i=1}^{N} v_i = 1$. Consequently, the PageRank vector P is the steady state vector, $P = GP$, of the Markovian process describing the endless journey of the random surfer. Furthermore, the PageRank vector is the right eigenvector of the Google matrix G associated with the eigenvalue 1. The ith component of the PageRank vector P gives the probability of finding the random surfer at node i after its endless journey, or, in practical terms, after a sufficiently high number of iterations in order to ensure the convergence toward the PageRank vector P. We order all the nodes initially labeled by $i = 1, \ldots, N$ by decreasing PageRank probability $\{P_j\}_{j=1,\ldots,N}$ introducing PageRank indexes $K = 1, \ldots, N$. A node j is ranked higher than a node i, i.e. $K(j) < K(i)$, if $P_j > P_i$. The index $K = 1$ is then assigned to the node with the highest PageRank probability, the index $K = 2$ is assigned to the node with the second highest PageRank probability, and so on. The index $K = N$ is assigned to the node with the lowest PageRank probability. The numerical computation of PageRank probabilities is performed efficiently using the PageRank iteration algorithm described in [8, 9].

It is also useful to consider the network obtained by reversing the direction of all directed links of the original network. With this reversed network, we associate the Google matrix G^*, which is built from the adjacency matrix $A^* = A^{\mathrm{t}}$ which is the transpose of the original adjacency matrix A. The elements of the stochastic matrix S^* are either $S^*_{ij} = A^*_{ij}/k^*_{out}(j) = A_{ji}/k_{in}(j)$, if $k_{in}(j) = k^*_{out}(j) = \sum_{k=1}^{N} A_{jk} \neq 0$, or $S^*_{ij} = 0$, otherwise. The elements of the Google matrix G^* are $G^*_{ij} = \alpha S^*_{ij} + (1-\alpha)/N$. The PageRank vector associated with G^* is called the CheiRank vector P^*. It characterizes the steady state, $P^* = G^* P^*$, of the Markovian process described by the matrix G^* [28, 29] (see also [10]). Its components, $\{P^*_i\}_{i=1;\ldots,N}$, are also probabilities that can again be ordered by decreasing values. We introduce the CheiRank indexes

$K^* = 1, \ldots, N$ in such a way that we assign $K^* = 1$ ($K^* = N$) to the node with the highest (lowest) CheiRank probability. On average, high PageRank (CheiRank) probabilities correspond to nodes with many ingoing (outgoing) links [9,10].

The PageRank measure of centrality characterizes the influence of a node inside the network since it quantifies how other nodes point toward it. By contrast, the CheiRank measure of centrality characterizes the diffusivity of a node since it quantifies how it points toward the other nodes of the network. To summarize, according to the PageRank algorithm, a node is all the more influential if it is pointed to by influential nodes while, according to the CheiRank algorithm, a node is all the more diffusive if it points to diffusive nodes.

Reduced Google matrix

The REGOMAX algorithm was proposed in [13] and is described in detail in [14–16,18]. Here, we present the main elements of this method using the notation employed in [14].

Let us consider a subset $\mathcal{S}_r$ of N_r nodes of interest from a huge network of $N \gg N_r$ nodes. The REGOMAX algorithm efficiently computes a *reduced Google matrix* G_{R} of size $N_r \times N_r$ that captures the full contributions of direct and indirect pathways between any pair of nodes among the N_r nodes of interest. For a random surfer navigating the entire network during an infinite journey, the reduced Google matrix entry $G_{\mathrm{R}ij}$ gives the probability of transition from node j to node i ($i, j \in \mathcal{S}_r$), while taking into account the probability of a possible direct transition from node j to node i, and also taking account of the indirect transition probabilities for all the indirect pathways from node j to node i via at least one node k not belonging to the nodes of interest subset $\mathcal{S}_r$. Mathematically, the reduced Google matrix is unequivocally determined once we assume that the relative PageRank probabilities of the N_r nodes obtained from the $N_r \times N_r$ reduced Google matrix G_{R} are the same as those obtained from the global $N \times N$ Google matrix G. The reduced Google matrix G_{R} can be decomposed into three matrix components that clearly distinguish between direct and indirect interactions, $G_{\mathrm{R}} = G_{rr} + G_{\mathrm{pr}} + G_{\mathrm{qr}}$ [14]. Here, the G_{rr} matrix component is a $N_r \times N_r$ submatrix of the global $N \times N$ Google matrix G corresponding to the direct transition probabilities between the N_r nodes of interest. The G_{pr} matrix component is quite similar to the matrix in which each column is equal to the PageRank vector P_r, which is defined as $G_{\mathrm{R}} P_r = P_r$. As the PageRank vector components $\{P_{ri} = \kappa P_i\}_{i \in \mathcal{S}_r}$ obtained from G_{R} are straightforwardly deduced from the PageRank vector components $\{P_i\}_{i \in \mathcal{S}_r}$, with the coefficient κ being a constant for all i, the G_{pr} matrix component does not provide much more information than the global PageRank vector P about direct and indirect links between the N_r selected nodes. Thus the most interesting and nontrivial role is played by the G_{qr} matrix component. This matrix component takes account of all indirect links between the N_r selected nodes emerging from myriads of pathways passing through multiple nodes of the global network but not belonging to the nodes of interest subset $\mathcal{S}_r$ (see [13,14]). We decompose the $G_{\mathrm{qr}} = G_{\mathrm{qrd}} + G_{\mathrm{qrnd}}$ matrix component into its diagonal part (G_{qrd}) and its non-diagonal part (G_{qrnd}). The G_{qrd} diagonal matrix component contains self-interactions of nodes which are not of interest for the present study. The precise formulas used to compute the three matrix components of the reduced Google matrix G_{R} are given in [13,14].

A useful additional characteristic provided by the G_{R} matrix is the sensitivity of the PageRank probability to the variation of a specific link between a pair of nodes chosen among the N_r nodes of interest. The useful results obtained with this method have been demonstrated in [15–17,19,20]. The PageRank sensitivity $D(j \to k, i)$ of a node i

to the matrix element $G_{\mathrm{R}jk}$ is

$$D(j \to k, i) \equiv \frac{1}{P_i} \frac{dP_i}{d\delta_{kj}} = \frac{1}{P_i(0)} \lim_{\delta_{kj} \to 0} \frac{P_i(\delta_{kj}) - P_i(0)}{\delta_{kj}}. \quad (1)$$

Here, $P(\delta_{kj})$ is the PageRank vector obtained from the reduced Google matrix G_{R}, where all the elements of jth column are modified as $\{G_{\mathrm{R}lj} \to G_{\mathrm{R}lj}/(1+\delta_{kj}G_{\mathrm{R}kj})\}_{l \neq k}$, with the exception of the kth element which is transformed as $G_{\mathrm{R}kj} \to G_{\mathrm{R}kj}(1+\delta_{kj})/(1+\delta_{kj}G_{\mathrm{R}kj})$. This transformation ensures that the sum of these transformed elements stays equal to 1, $\sum_{l=1}^{N_r} G_{\mathrm{R}lj} = 1$. This is a mandatory property since the elements of the jth column of G_{R} represent the probabilities of transitions from node j to the other nodes. Roughly speaking, forgetting the denominator $(1+\delta_{kj}G_{\mathrm{R}kj})$, which ensures the normalization of the jth column, the PageRank sensitivity $D(j \to k, i)$ measures the relative variation of the PageRank of the ith node when the G_{R} link from node j to node k changes by an infinitesimal amount.

Table 1. List of the 34 largest pharmaceutical companies ranked by the relative PageRank index K_r of their corresponding articles in Wikipedia. The K_{LMC} index gives the ranking by the largest market capitalization since 2000 [2] and the K_{MC} index gives the ranking by market capitalization in 2017 [2]. The relative CheiRank index K_r^* is also given.

K_r	K_r^*	K_{LMC}	K_{MC}	Company	K_r	K_r^*	K_{LMC}	K_{MC}	Company
1	3	2	3	Pfizer	18	16	18	18	Biogen
2	1	8	15	GSK[a]	19	25	28	28	Mylan
3	8	14	10	Bayer	20	13	7	13	Gilead
4	10	1	1	J&J[b]	21	17	21	21	Shire
5	2	3	4	Novartis	22	22	23	23	Takeda
6	4	6	6	Merck	23	28	29	27	Astellas
7	6	16	14	Lilly	24	18	32	29	Daiichi Sankyo
8	11	4	2	Roche	25	15	5	5	AbbVie
9	5	19	16	AstraZeneca	26	32	22	22	Regeneron
10	12	10	9	Sanofi	27	24	33	31	Eisai
11	9	15	11	BMS[c]	28	31	20	19	Stryker
12	7	13	12	Abbott	29	19	34	33	BioMarin
13	29	24	24	Illumina	30	26	27	26	Zoetis
14	14	11	8	Amgen	31	23	25	25	Vertex
15	21	9	7	Novo Nordisk	32	30	26	30	Alexion
16	33	12	20	Allergan	33	34	31	34	Perrigo
17	20	17	17	Celgene	34	27	30	32	Incyte

[a]GSK: GlaxoSmithKline, [b]J&J: Johnson & Johnson, [c]BMS: Bristol-Myers Squibb.

Datasets

We consider the English Wikipedia edition as at May 2017 with $N = 5\,416\,537$ articles (nodes) and $N_l = 122\,232\,932$ hyperlinks between articles. This network has also been considered in [16, 17, 19, 20]. For the REGOMAX analysis, we select $N_c = 195$ world countries (see the list and PageRank order in [19, 20]), the $N_{ph} = 34$ largest pharmaceutical companies (see Table 1), $N_{rd} = 47$ rare renal diseases (see Table 2), and $N_{cr} = 37$ types of cancer listed in [20]. Thus we consider a total of $N_r = N_c + N_{ph} + N_{rd} + N_{cr} = 313$ Wikipedia articles as nodes of interest. All data sets are available at [30]. The present study complies with the Wikimedia terms of use.

Table 2. List of 47 rare renal diseases ranked by the relative PageRank index K_r of their corresponding articles in Wikipedia. The relative CheiRank index, K_r^*, is also given. The list of rare renal diseases is split in 5 categories with the following color code: ■ Congenital abnormalities of the kidney and urinary tract, ■ Glomerular diseases, ■ Renal tubular diseases and metabolic diseases, ■ Nephrolithiasis and ■ Ciliopathies and Nephronophthisis.

K_r	K_r^*		Rare renal disease	Short name
1	6		Alport syndrome	Alport
2	1		Bardet–Biedl syndrome	Bardet–Biedl
3	7		Fabry disease	Fabry
4	2		Kallmann syndrome	KS
5	4		Renal tubular acidosis	RTA
6	5		Cystinuria	Cystinuria
7	14		Renal agenesis	Renal agenesis
8	16		Nephronophthisis	Nephronophthisis
9	33		X-linked hypophosphatemia	XLH
10	8		Bartter syndrome	Bartter
11	41		Xanthinuria	Xanthinuria
12	3		Oculocerebrorenal syndrome	Oculocerebrorenal
13	17		Nail–patella syndrome	Nail-patella
14	45		Familial renal amyloidosis	Amyloidosis
15	9		Episodic ataxia	EA
16	24		Gitelman syndrome	Gitelman
17	21		Medullary cystic kidney disease	MCKD
18	32		Familial hypocalciuric hypercalcemia	FHH
19	42		Renal glycosuria	Glycosuria
20	40		Fanconi-Bickel syndrome	Fanconi-Bickel
21	23		Liddle's syndrome	Liddle
22	31		Hypomagnesemia with secondary hypocalcemia	HSH
23	13		Fraser syndrome	Fraser
24	36		WAGR syndrome	WAGR
25	19		Branchio-oto-renal syndrome	BOR
26	26		Townes–Brocks syndrome	Townes–Brocks
27	44		Autosomal dominant hypophosphatemic rickets	ADHR
28	18		Orofaciodigital syndrome 1	Orofaciodigital
29	28		Denys–Drash syndrome	DDS
30	10		Perlman syndrome	Perlman
31	11		Simpson–Golabi–Behmel syndrome	SGBS
32	12		Caroli disease	Caroli
33	25		Congenital nephrotic syndrome	GNS
34	39		Fechtner syndrome	Fechtner
35	27		Abderhalden–Kaufmann–Lignac syndrome	AKL
36	22		Lysinuric protein intolerance	LPI
37	47		Majewski's polydactyly syndrome	Majewski
38	20		Urofacial syndrome	Urofacial
39	15		Barakat syndrome	Barakat
40	35		Dicarboxylic aminoaciduria	DCBXA
41	38		MODY 5	MODY 5
42	30		Papillorenal syndrome	Papillorenal
43	43		OCRL	OCRL
44	37		COQ9	COQ9
45	29		Juvenile nephronophthisis	J. nephronophthisis
46	46		Renal-hepatic-pancreatic dysplasia	RHPD
47	34		Conorenal syndrome	Conorenal

In Table 2, the 47 rare renal diseases can be grouped into five categories with which we have associated a color to simplify identification: congenital abnormalities of the kidney and urinary tract (red), glomerular diseases (blue), renal tubular diseases and metabolic diseases (gold), nephrolithiasis (cyan), and ciliopathies and nephronophthisis (green). This reflects the categorization used in [27]. Of the 166 genetic disorders of renal growth and of renal function listed in [27], only the 47 rare renal diseases cited above had a corresponding article in Wikipedia dated at May 2017.

Results

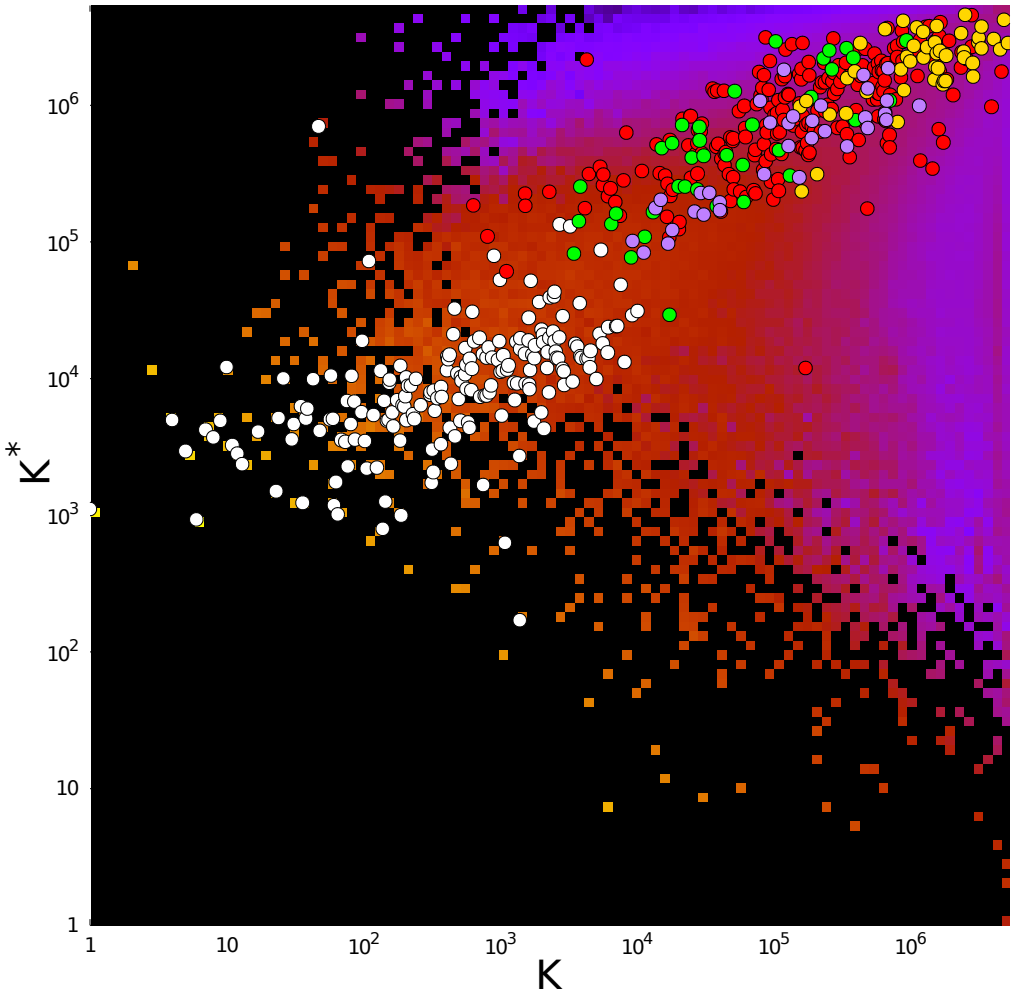

Fig 1. Density of Wikipedia articles in the PageRank-CheiRank plane (K, K^*). Data are averaged over a 100×100 grid for $(\log_{10} K, \log_{10} K^*)$ spanning the domain $[0, \log_{10} N] \times [0, \log_{10} N]$. Density of articles ranges from very low (purple tiles) to very high (bright yellow tiles). The absence of any article is represented by black tiles. The superimposed white circles give the positions of articles devoted to the 195 countries, the red circles represent the positions of articles devoted to the 230 kinds of infectious diseases studied in [19], the green circles represent the positions of 37 articles on the cancers studied in [20], the gold circles mark the positions of 47 articles on rare renal diseases and the purple circles give the positions of 34 articles on pharmaceutical companies.

PageRank vs CheiRank distributions

In Fig 1 we show the coarse-grained distribution of all N Wikipedia articles on the PageRank-CheiRank plane (K, K^*) in logarithmic scale. On this plane, we plot the positions of 195 countries (white circles), 37 cancer types (green circles), 47 rare renal diseases (gold circles), 34 pharmaceutical companies (purple circles) and we also add the 230 infectious diseases (red circles) studied in [19]. As usual, countries take the top PageRank positions (see also [10]), followed by the groups of infectious diseases and cancers. The group of rare diseases is clustered in the high K and high K^* region, being generally far lower-ranked. The highest-ranked company in the group of pharmaceutical companies is a little behind the leader of the group of cancers. In the global PageRank list the top 3 in each group are $K = 1$ US, $K = 2$ France, $K = 3$ Germany for countries; $K = 639$ tuberculosis, $K = 810$ HIV/AIDS, $K = 1116$ malaria for infectious diseases [19]; $K = 3478$ lung cancer, $K = 3788$ breast cancer, $K = 3871$ leukemia for cancers [20]; $K = 9345$ Pfizer, $K = 11290$ GlaxoSmithKline, $K = 13737$ Bayer for pharmaceutical companies, $K = 156963$ Alport syndrome, $K = 161731$ Bardet–Biedl syndrome, $K = 174780$ Fabry disease for rare renal diseases.

In Table 1, we give the relative PageRank K_r list of 34 pharmaceutical companies. In this table, we also indicate the rank K_{LMC} of these companies in terms of largest market capitalization [2] and the rank K_{MC} based on their market capitalization ranking in 2017 [2]. The CheiRank rank K_r^* is also given. To permit a quick comparison, Table 3 presents the Spearman rank correlation coefficients ρ between these rankings. As expected, the K_{LMC} and K_{MC} rankings are similar being highly correlated. We observe that the PageRank K_r and CheiRank K_r^* are also highly correlated. Consequently, on average, the influence and the diffusivity of an article on a pharmaceutical company in Wikipedia are correlated. We also observe a fairly high correlation ~ 0.8 between PageRank K_r and market capitalization rankings K_{MC} or K_{LMC}. The influence of pharmaceutical company articles within Wikipedia is consequently related to the real economic performances of the pharmaceutical companies. The correlation between CheiRank K_r^* and the K_{MC} or K_{LMC} rankings is a little weaker at about ~ 0.7.

The distribution of the companies on the rank plane (K_r, K_{LMC}) is shown in Fig 2. If the influence in Wikipedia of the pharmaceutical companies were strictly proportional to their wealth, the companies would lie on the diagonal, $K_r = K_{LMC}$. Consequently, in this figure, companies above (below) the diagonal have an excess (shortfall) of influence in Wikipedia relative to their market capitalization. The richest company, Johnson & Johnson (J&J), is only at the 4th position in PageRank ($K_r = 4$). This shows that its influence in Wikipedia is well behind its first place in terms of market capitalization. We can also note that AbbVie, a spin-off of Abbott laboratories, which has enjoyed the 5th largest market capitalization since 2000 is among the less influential of the considered pharmaceutical companies with $K_r = 25$. In contrast, Pfizer is the most influential company on Wikipedia with $K_r = 1$ and $K_{LMC} = 2$. A significant influence in Wikipedia is found for pharmaceutical companies with a leading K_r

Table 3. Spearman rank correlation coefficients ρ between the K_r, K_r^*, K_{LMC}, and K_{MC} rankings for pharmaceutical companies presented in Table 1.

ρ	K_r	K_r^*	K_{LMC}	K_{MC}
K_r	1.00	0.80	0.76	0.79
K_r^*		1.00	0.67	0.73
K_{LMC}			1.00	0.96
K_{MC}				1.00

position and a somewhat worse K_{LMC}. Such companies include GlaxoSmithKline (GSK) with $K_r = 2$ and $K_{LMC} = 8$ and Bayer with $K_r = 3$ and $K_{LMC} = 14$.

The overlap $\eta(j)$ between the two ranking lists K_r and K_{LMC} (see Table 1), as well as between the two ranking lists K_r and K_{MC} (see Table 1), is shown in Fig 3. We see that for the first three positions, the overlap is only $\eta = 1/3$. Assuming that the major search engines still probe the structural linking properties of the WWW, this shows that the richest pharmaceutical companies can significantly improve their public visibility in Wikipedia. At present, the Wikipedia web site is the 5th most frequently visited site in the world [31] and the improvement in the content of Wikipedia articles can bring significantly better global visibility for companies; furthermore article editing is free of charge. Nevertheless 7 of the top 10 most influential pharmaceutical companies in Wikipedia are also among the top 10 in terms of market capitalization since 2000 and among the top 10 for market capitalization in 2017.

We have to mention that, on average, older Wikipedia articles, i.e., articles with earlier creation dates, may have accumulated more ingoing and outgoing links than newer ones and that, consequently, older Wikipedia articles should, on average, have better PageRank and CheiRank indexes. This effect is not observed among articles devoted to countries, as almost all of these articles were created during the same period, i.e., in 2001 at the very beginning of Wikipedia. However, on average, this effect is present, albeit with considerable fluctuations, among articles devoted to cancer types and pharmaceutical companies.

The distribution of the 47 articles on rare renal diseases on the PageRank-CheiRank plane is shown in Fig 4. The top three PageRank positions are taken by Alport syndrome, Bardet-Biedl syndrome and Fabry disease ($K_r = 1, 2, 3$), meaning that these are most influential rare renal diseases. The top three CheiRank positions are held by Bardet-Biedl syndrome, Kallmann syndrome and Oculocerebrorenal syndrome ($K_r^* = 1, 2, 3$), making them the most diffusive articles among the 47 articles devoted to rare renal diseases. Subsets of articles in Wikipedia such as, e.g., articles devoted to Universities [12], to infectious diseases [19] or to cancer types [20], are usually distributed around the $K_r = K_r^*$ diagonal. We note that the case of rare renal diseases is different since, with the exception of the top 6 diseases in K_r and a few others, the rare renal diseases are located far from the $K_r = K_r^*$ diagonal. This is corroborated by the Spearman rank correlation coefficient of $\rho \simeq 0.54$. The diseases above (below) the diagonal are more influential (diffusive) than diffusive (influential) in the Wikipedia network. We have to remember that these K_r and K_r^* rankings are relative to the considered subset of renal diseases with high K and K^* indexes, and that none of the articles devoted to these diseases are either influential or diffusive in the global Wikipedia network. Hence, for rare renal diseases, high K_r – low K_r^* articles have more outgoing links than the average in this subset pointing toward mildly diffusive articles, e.g., articles containing lists of rare diseases. High K_r^* – low K_r articles have more ingoing links than the average in this subset pointing from mildly influential articles, e.g., the top K_r articles on rare renal diseases.

Example of a reduced Google matrix G_{R}

We will consider Wikipedia articles (nodes of interest) among the 313 selected articles (nodes) devoted to countries ($N_c = 195$), pharmaceutical companies ($N_{ph} = 34$), cancers ($N_{cr} = 37$) and rare renal diseases ($N_{rd} = 47$). Since the interactions between countries and cancers have been analyzed in [20], we show other interactions below, such as those between pharmaceutical companies and countries, cancers, rare renal diseases, and between rare renal diseases and countries. In Fig 5, we show the 81×81 reduced Google matrix G_{R} (top left panel) and its G_{rr} (top right panel), G_{pr} (bottom left panel), G_{qrnd} (bottom right panel) components associated with pharmaceutical companies and rare

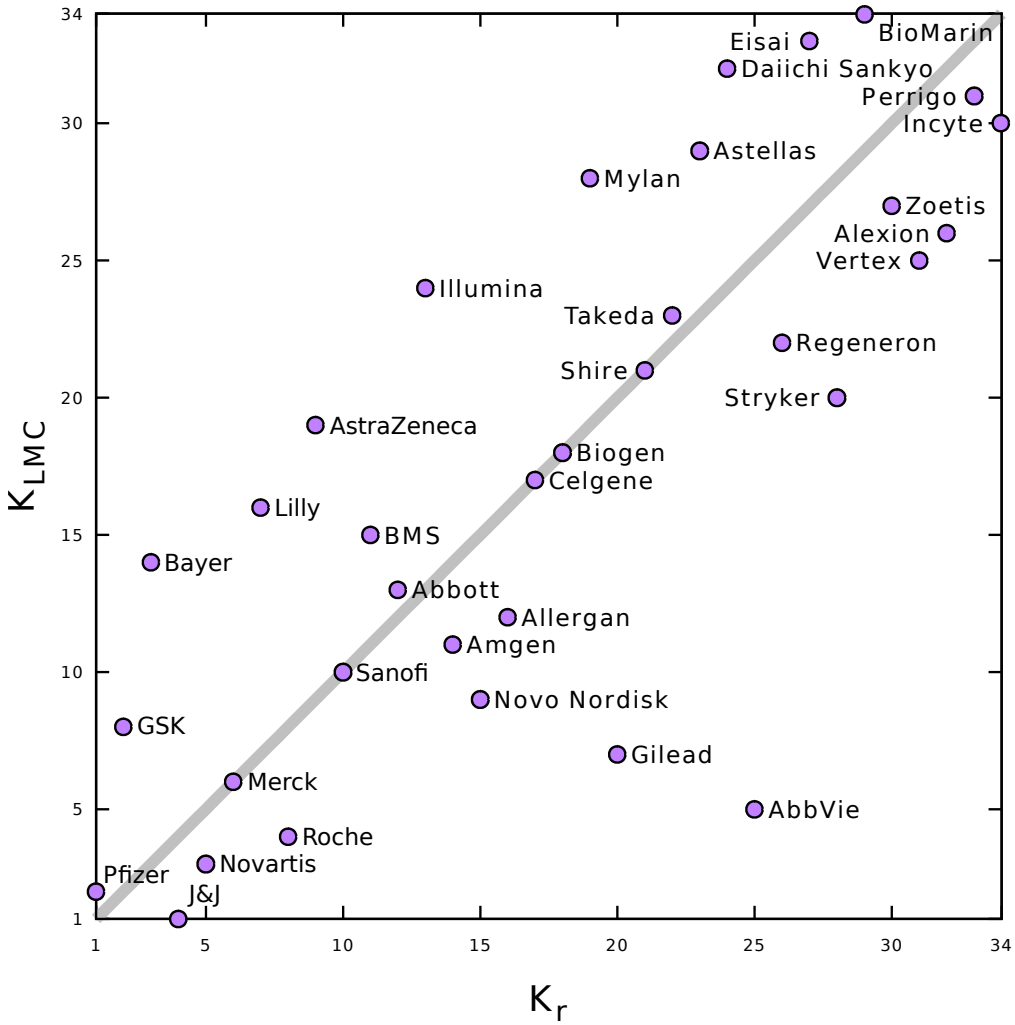

Fig 2. Distribution of pharmaceutical companies ranked by largest market capitalization index, K_{LMC}, and by the relative PageRank index, K_r, of their article in Wikipedia. See rankings in Table 1.

renal diseases. Focusing on G_{R} (Fig 5 top left panel), the first 34×34 block diagonal sub-matrix (delimited by the purple square) gives the effective directed links between pharmaceutical companies. The remaining 47×47 block diagonal sub-matrix (subdivided into five block diagonal sub-matrices delimited by colored squares; the color codes for the rare renal diseases are indicated in Table 2) shows the effective directed links between the rare renal diseases. The 47×34 (34×47) rectangular sub-matrix indicates the effective directed links from (to) pharmaceutical companies to (from) rare renal diseases. We clearly see from the G_{pr} component (Fig 5 bottom left panel) that the columns are very similar, since mathematically each column is approximately equal to the PageRank vector P_r. Consequently, the G_{pr} components in their turn provides us with already known information about the global PageRank. The almost uniform rainbow pattern is due to the fact that the columns are similar and the entries are ordered by decreasing PageRank probabilities. We retrieve the fact that Wikipedia articles about top pharmaceutical companies have a higher PageRank than those concerning rare renal diseases. The 22 most influential companies have a global PageRank index from $K \sim 10^4$ to $K \sim 10^5$ (see Fig 1), whereas the most influential

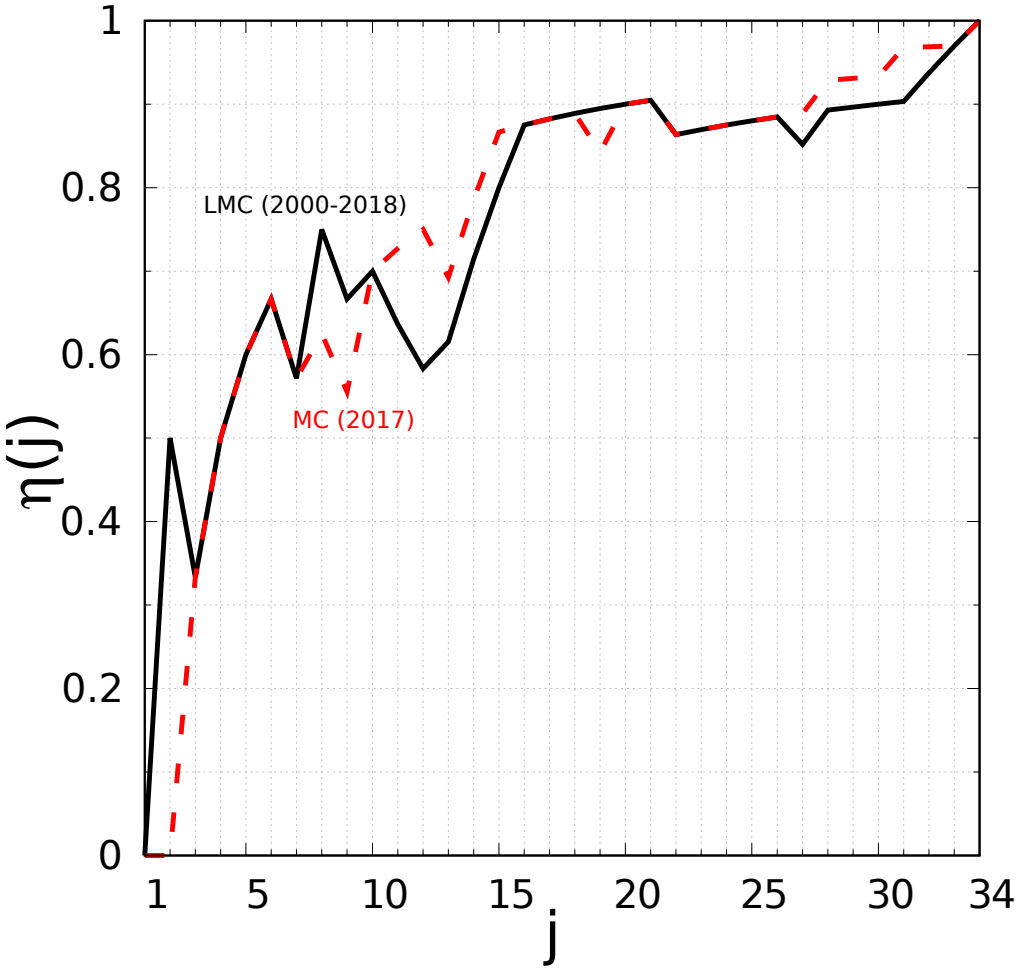

Fig 3. Overlap between the PageRanking of articles on pharmaceutical companies in Wikipedia and the ranking of pharmaceutical companies by market capitalization. The overlap function is $\eta(j) = j_{ph}/j$, where j_{ph} is the number of pharmaceutical companies present in the top j of both lists: the PageRanking of pharmaceutical companies articles in Wikipedia (see first column of Table 1) and the list of pharmaceutical companies ranked by market capitalization in 2017 (red curve, see third column in Table 1) and ranked in terms of the largest market capitalization since 2000 (black curve, see second column in Table 1).

rare renal disease have a PageRank index above $K \sim 10^5$. This is also clearly seen in the $G_{\rm pr}$ matrix picture (Fig 5, bottom left panel). The direct links are visible in the G_{rr} matrix (Fig 5 top right panel), which provides us with a picture of the adjacency matrix A. We observe that direct links are denser in the block diagonal for pharmaceutical companies than in that for rare renal diseases. In the pharmaceutical companies sector there are many direct links, with almost every Wikipedia article devoted to a pharmaceutical company pointing toward at least one other (with the exception of Alexion, BioMarin and Novo Nordisk). This is due to the very competitive nature of the industry, in which it is routine for companies to acquire drug departments from one another. Still looking at the top right panel of Fig 5, we observe that in the rare renal diseases sector, the few direct links concern rare renal diseases belonging to the same category (pixels inside the red, blue, gold cyan and green squares). The exceptions are

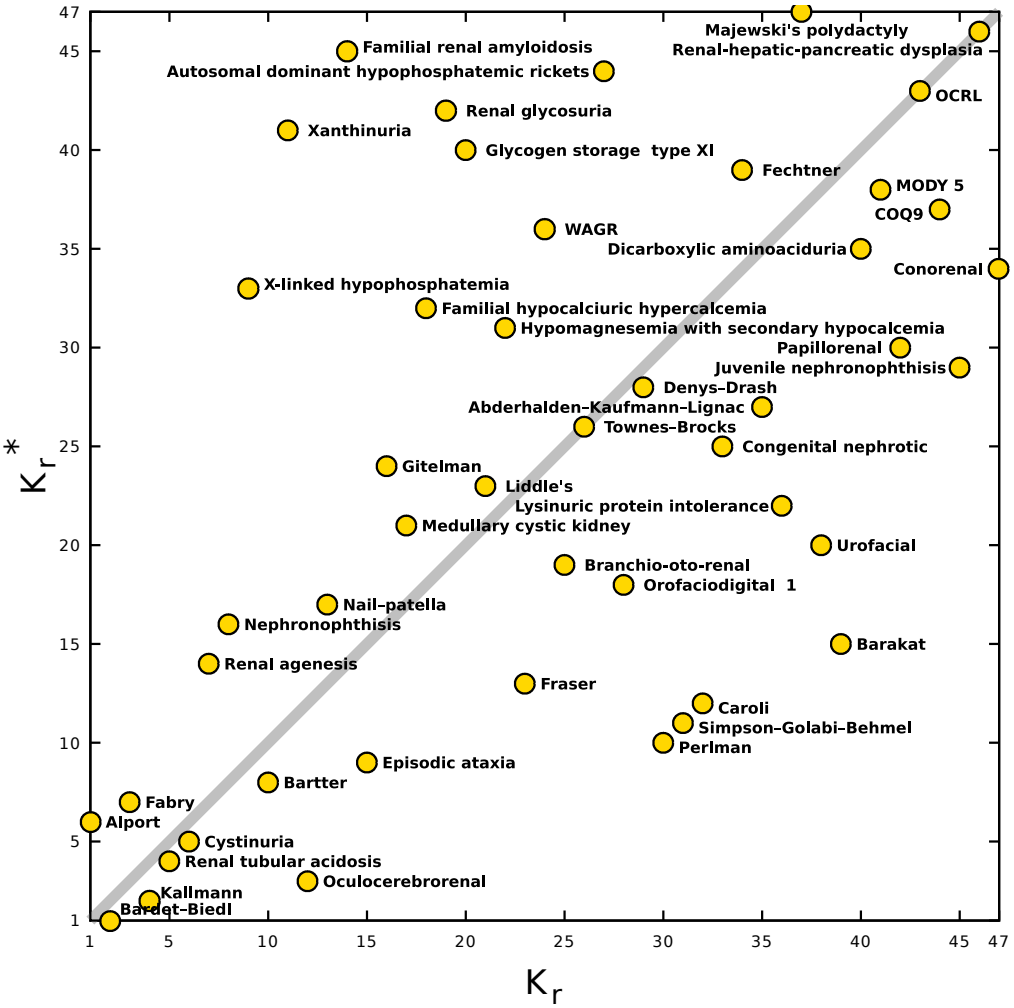

Fig 4. Distribution of $N_{rd} = 47$ articles on rare renal diseases on the plane of relative PageRank-CheiRank indexes (K_r, K_r^*); positions in the plane are indicated by golden circles with short names for the diseases.

nephronophthisis, which belongs to the ciliopathies category, and medullary cystic kidney disease, which belongs to the renal tubular diseases and metabolic diseases category; each one these diseases points directly to the other. In fact, these two diseases share similar morphological and clinical features (see, e.g., [32]), thus explaining the reciprocal direct links between them. We also observe that none of the Wikipedia articles devoted to the pharmaceutical companies with the largest market capitalization directly cite one of the rare renal diseases. This observation highlights the orphan status of these diseases. The only direct link existing between pharmaceutical companies and rare renal diseases is observed in connection with Fabry disease, which points toward Shire the manufacturer of Replagal, a dedicated drug for Fabry disease [33]. Finally, the $G_{\rm qr}$ matrix component (Fig 5 bottom right panel) indicates indirect links, some of which are purely hidden links since they do not appear in the G_{rr} matrix component (Fig 5 top right panel). The $G_{\rm qr}$ matrix component reveals many hidden links between pharmaceutical companies which are certainly due to the complex economic interconnections between these companies. The two strongest hidden links appear from Alexion to Shire and from Abbott to Johnson & Johnson. Many indirect pathways

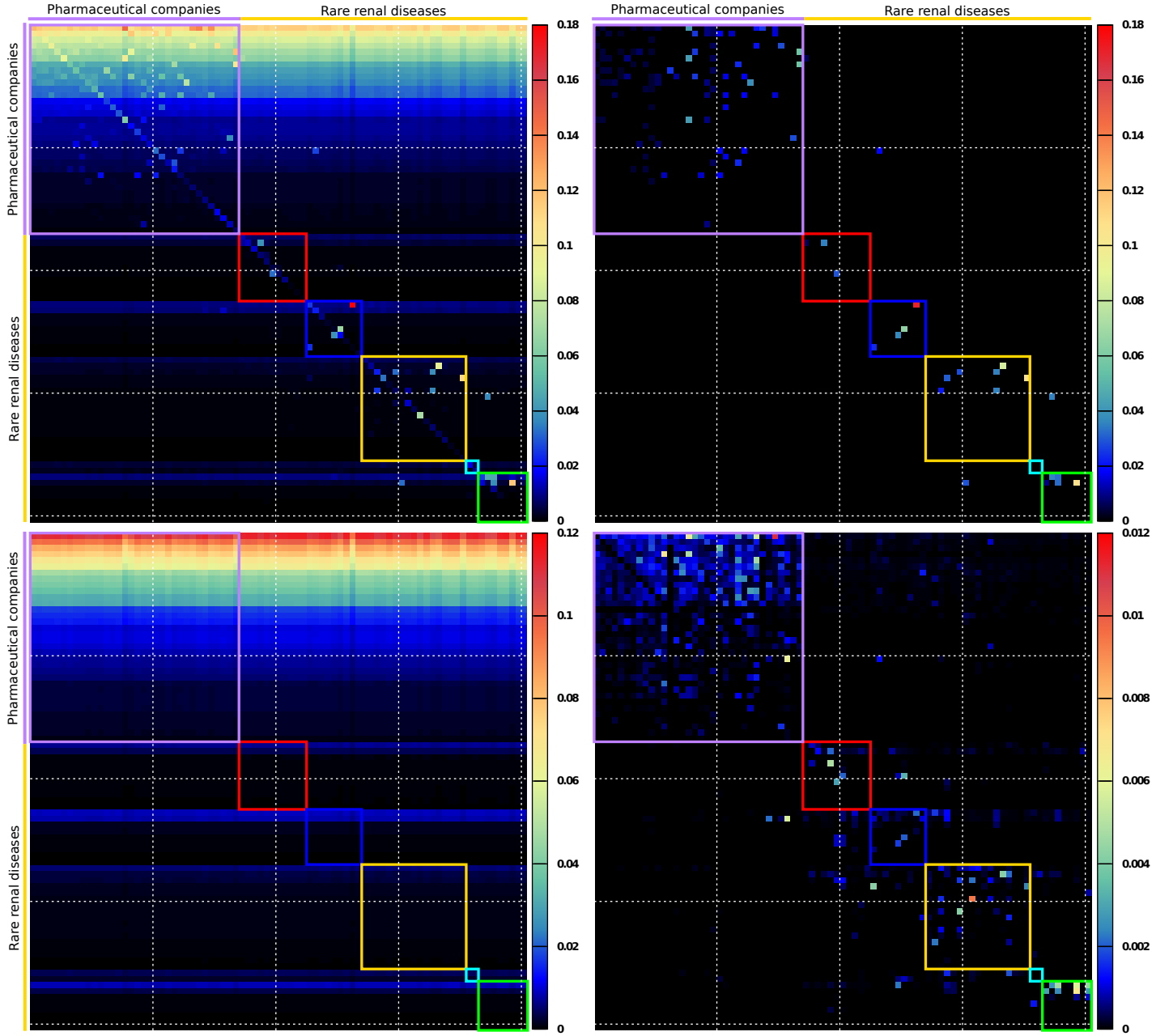

Fig 5. Reduced Google matrix G_{R} of pharmaceutical companies and rare renal diseases. We show the reduced Google matrix G_{R} (top left panel) and its three components G_{rr} (top right panel), G_{pr} (bottom left panel), and G_{qrnd} (bottom right panel). Each "pixel" represents a matrix entry with the amplitude given by a color. Color bars give the corresponds between matrix entry amplitudes and colors. For each 81×81 matrix the first 34 entries correspond to pharmaceutical companies (ordered as in Table 1) and the other 47 entries correspond to rare renal diseases (ordered by categories then by PageRank order inside each category, see Table 2). The first 34×34 block diagonal sub-matrix (purple square) corresponds to directed interactions between pharmaceutical companies. The other five smallest block diagonal sub-matrices correspond to directed interactions between rare renal diseases belonging to one of the five categories defined in Table 2. The colors of the squares correspond to color categories given in Table 2. For the sake of visibility horizontal and vertical white dashed lines are drawn after every 20 entries.

through the overall Wikipedia network contribute to these hidden links. E.g., one of the shortest paths going from Alexion to Shire is the Alexion → Ludwig N. Hantson → Shire link [34]. Indeed, Ludwig N. Hantson was appointed as CEO of Alexion Pharmaceuticals in March 2017 [35], and he was CEO of Baxalta prior its acquisition by Shire in June 2016 [36]. Furthermore, the indirect link Abbott → Advanced Medical Optics → Johnson & Johnson contributes to the Abbott → Johnson & Johnson hidden link. Indeed, Advanced Medical Optics, also known as Abbott Medical Optics, was acquired by Johnson & Johnson in February 2017 [37]. Concerning the rare renal diseases sector, the strongest hidden links among diseases belonging to the same category are: Familial hypocalciuric hypercalcemia → Gitelman syndrome, as both are hypocalciuric diseases [38]; the inverted link Gitelman syndrome → Familial hypocalciuric hypercalcemia, Juvenile nephronophthisis → Bardet–Biedl syndrome, as nephronophthisis and Bardet–Biedl syndrome are both typically cited examples of ciliopathies; Townes–Brocks syndrome → Branchio-oto-renal syndrome, as both of these syndromes imply sensorineural hearing loss [39]. Between diseases belonging to different categories, we observe the notable hidden links: Alport syndrome → Oculocerebrorenal syndrome (the inverted hidden link is also present but weaker), as both are examples of genetic causes of cataracts in childhood or early life [40]; Denys–Drash syndrome → Perlman syndrome, since both lead to a high risk of Wilms' tumor [41,42]. The most interesting hidden links are those that occur between pharmaceutical companies and rare renal diseases. The strongest of these is Alexion → Fabry disease. Indeed, in May 2015, Alexion Pharmaceuticals acquired Synageva BioPharma Corp., a company dedicated to rare diseases and to Fabry disease in particular [43]. We also observe hidden links, although somewhat weaker, from BioMarin and Vertex to Fabry disease. Here, it is not easy to find indirect pathways through Wikipedia from Alexion, BioMarin, or Vertex to Fabry disease unlike the case for hidden links exclusively between the cited pharmaceutical companies or exclusively between rare renal diseases.

Reduced network structures

As the $G_{\rm pr}$ matrix component of the $N_r \times N_r$ reduced Google matrix $G_{\rm R}$ gives trivial results that are already known from a direct PageRank analysis of the $N \times N$ global Google matrix, we use the $G_{rr} + G_{\rm qr}$ matrix component to infer an effective reduced network between the N_r nodes of interest.

First, we build the reduced network of interactions between the $N_r = N_{ph} + N_c$ nodes of interest constituted by the $N_{ph} = 34$ pharmaceutical companies and the $N_c = 195$ countries (see Fig 6). The rules for the construction of the reduced network are described in the caption to Fig 6. At the first level, we take the top 5 companies in the PageRank list (Table 1) and trace the links with the largest column matrix elements in $G_{rr} + G_{\rm qr}$ from a company to two countries and two other companies. The process is then repeated to other levels from the newly added companies. The process is stopped when no new companies can be added. We obtain a compact reduced network of 15 pharmaceutical companies from the 34 initially selected, and 12 of these constitute the top 12 of the PageRank list (Table 1). The 3 remaining companies are nevertheless in the top 9 of pharmaceutical companies with the largest market capitalization since 2000.

By following direct links (black arrows) in this network, i.e., hyperlinks from Wikipedia articles, we see, e.g., that for Bayer, the closest companies are Sanofi and Roche and the two country friends are Germany, where its main office is located, and Belarus. These two links are direct links from the Bayer page in Wikipedia. Here, Belarus is cited as an example of a country from the Commonwealth of Independent States (CIS) in which the over-the-counter drug business was developing in early 2000s. To respond to this need, Bayer acquired Sagmel, Inc., a company already present in the CIS, in June 2008 [45,46]. For Pfizer, the two company friends are GlaxoSmithKline and

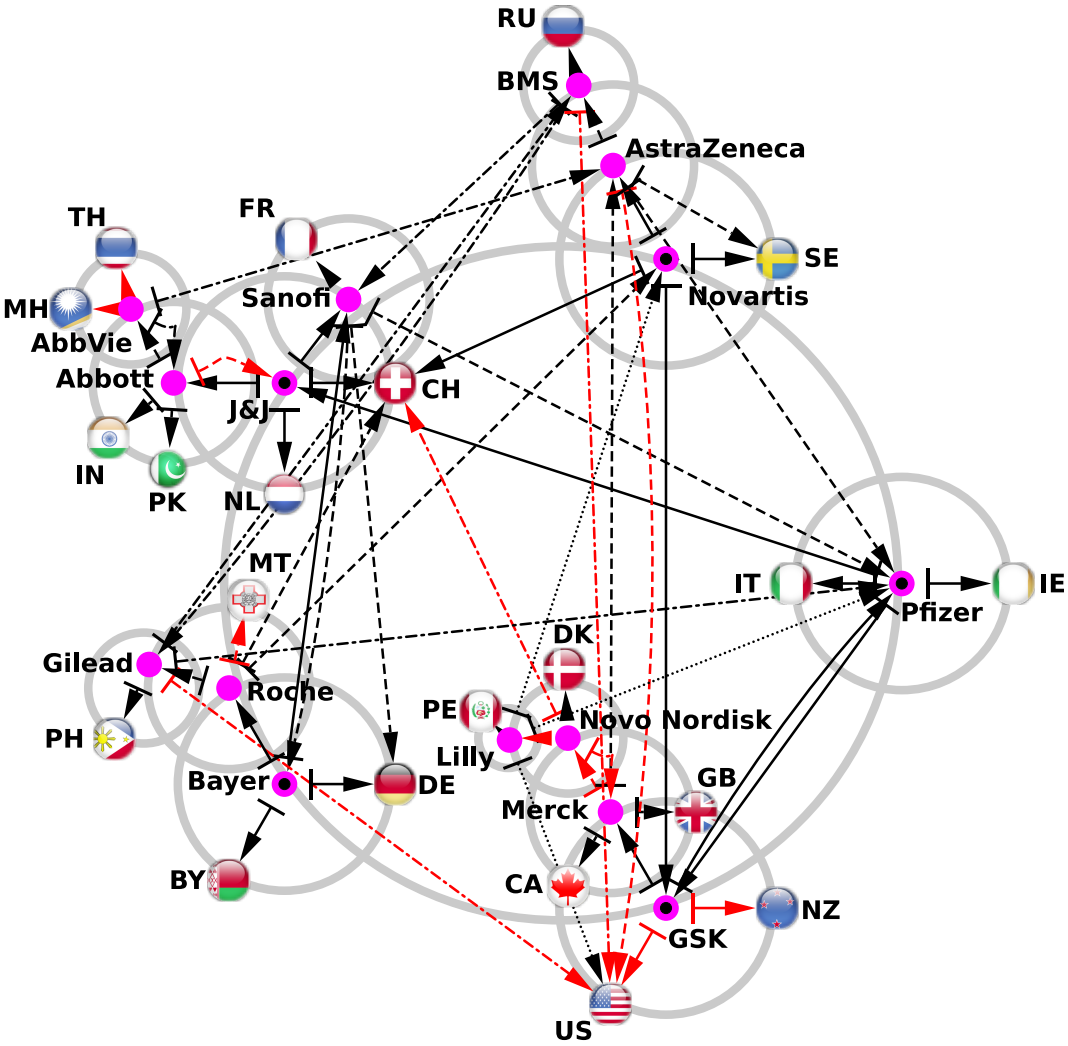

Fig 6. Reduced network of pharmaceutical companies with the addition of their best connected countries. We consider the first five pharmaceutical companies from the Wikipedia PageRank list: Pfizer, GSK, Bayer, J&J, and Novartis (see Table 1). Each one of these companies are represented by purple circles (●) placed along the main grey circle (the grey circle with the largest radius). From these five most influential pharmaceuticals in Wikipedia, we determine the two best connected companies, i.e., for a pharmaceutical company ph, we determine the two companies ph_1 and ph_2 giving the highest $(G_{rr} + G_{\rm qr})_{ph_{1\,\rm or\,2},ph}$ values. If not already present in the network, we add these best connected companies along secondary circles centered on the previous companies. The newly added companies are represented by purple circles (●). Also from the initial five pharmaceutical companies we determine the two best connected countries, i.e., for a company ph, we determine the two countries c_1 and c_2 giving the highest $(G_{rr} + G_{\rm qr})_{c_{1\,\rm or\,2},ph}$ values. From the newly added pharmaceutical of the first iteration, we determine the two best connected pharmaceutical companies and the two best connected countries. This constitutes the second iteration, an so on. At the fourth iteration of this process, no new pharmaceutical companies are added, and consequently the network construction process stops. The links obtained at the first iteration are represented by solid line arrows, at the second iteration by dashed line arrows, at the third iteration by dashed-dotted line arrows, and at the fourth iteration by dotted line arrows. Black color arrows correspond to links existing in the adjacency matrix (direct hyperlinks in Wikipedia), and red color arrows are purely hidden links absent from the adjacency matrix but present in $G_{\rm qr}$ component of the reduced Google matrix $G_{\rm R}$. The obtained network is drawn with the Cytoscape software [44]. Countries are marked by their ISO 3166-1 alpha-2 codes.

Johnson & Johnson, while the two country friends are Italy and Ireland. All these direct links are easy to explain by examining the hyperlinks in Wikipedia articles. Direct links between companies indicates major transactions between them that are reported in the dedicated Wikipedia articles. These direct links give a compact picture of the relationships between pharmaceutical companies. The non-obvious hidden links (red arrows in Fig 6) are potentially more interesting. Here, most of the hidden links point to the US, highlighting the fact that most of the pharmaceutical companies are American.

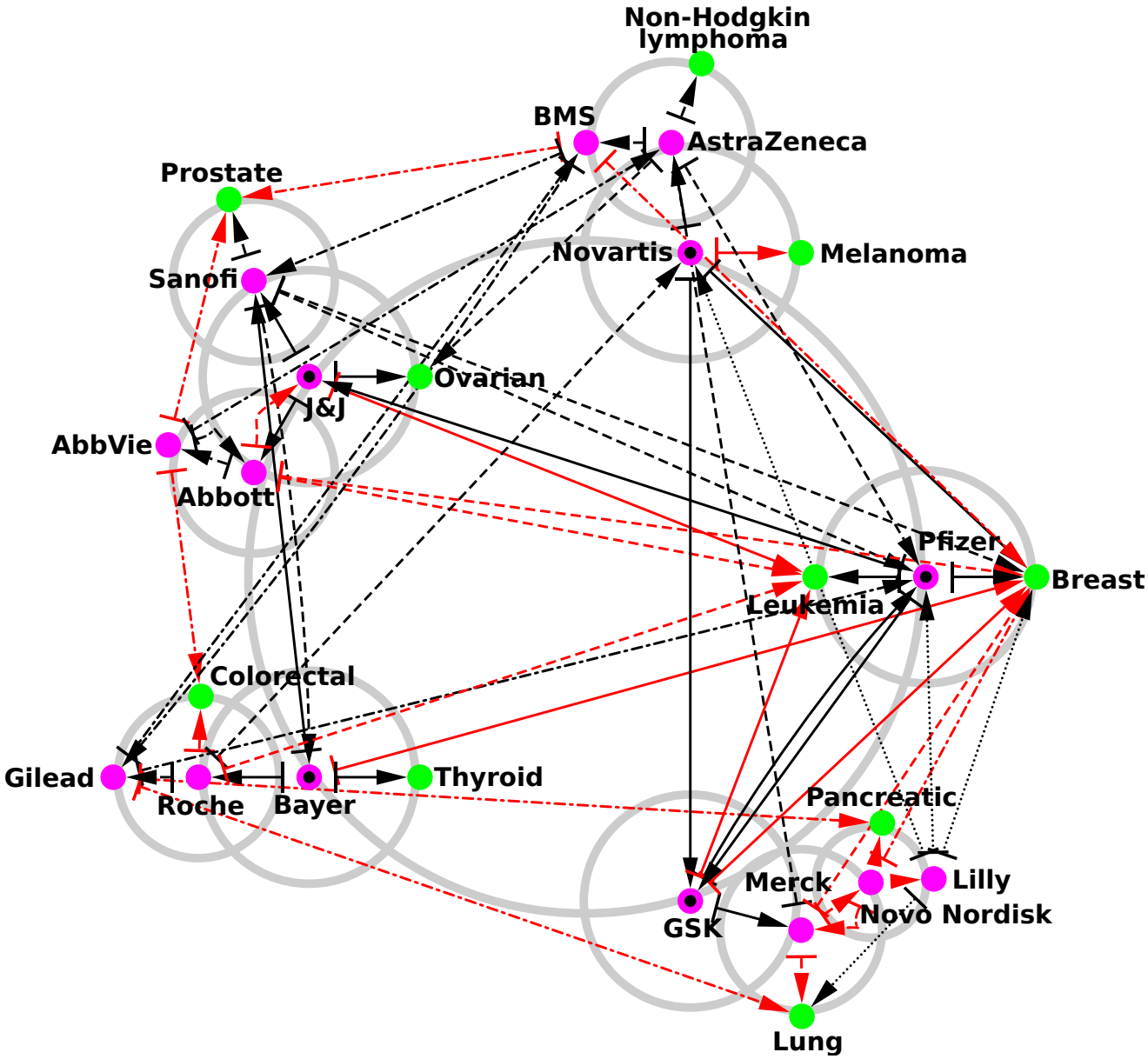

Fig 7. Reduced network of pharmaceutical companies with the addition of their best connected cancers. The construction algorithm is the same as the one used to generate Fig 6 excepting that we replace at each iteration the two best connected countries by the two best connected cancers. Pharmaceutical companies are represented by purple circles (●) and cancers by green circles (●).

The reduced network of interactions between the $N_{ph} = 34$ pharmaceutical companies and the $N_{cr} = 37$ cancer types is shown in Fig 7. At each level, it shows the

two closest company friends and two cancer types. The construction rules are the same as in Fig 6. The structure of the reduced network between pharmaceutical companies remains the same as in Fig 6. The reduced network indicates the strongest links from a company to the related types of cancer. We observe a clear polarization toward breast cancer, since 10 of the 15 pharmaceutical companies preferentially point to breast cancer, the Wikipedia article for which is the second most influential article among the 37 articles devoted to cancer types [20]. The second and third most connected cancer types are leukemia and lung cancer, which are preferentially pointed to by 4 and 3 pharmaceutical companies, respectively. Although the list (lung cancer, breast cancer, and leukemia) contains the 3 most influential cancer types in Wikipedia [20], we observe the particular interest in breast cancer on the part of pharmaceutical companies. More specifically, Pfizer and GlaxoSmithKline are most strongly linked to breast cancer and leukemia, while Bayer is most strongly linked to breast and thyroid cancers. In the case of Johnson & Johnson, the strongest links are to leukemia and ovarian cancers. For Novartis, the strongest links are with melanoma and breast cancers. Links to cancers for other companies are also clearly visible in Fig 7. We argue that the REGOMAX approach allows us to determine, from Wikipedia, the main orientations of pharmaceutical companies in their treatment of cancers. From Fig 7, we also see that in addition to direct links (black arrows), the indirect links (red arrows) also play a very important role.

The reduced network of pharmaceutical companies and rare renal diseases is shown in Fig 8. Its construction rules are the same as for Figs 6 & 7. Again, the network structure of the pharmaceutical companies remains the same as in Figs 6 & 7. Since 15 pharmaceutical companies are present, 30 preferentially connected rare renal diseases could have potentially emerged from this reduced network, but in fact only 10 rare renal diseases are present. From this network, we directly see the main orientations of each company toward specific rare renal diseases. We observe a strong polarization toward Fabry disease ($K_r = 3$ in Table 2) since 12 of the 15 pharmaceutical companies point to this disease. The second and third most connected rare renal diseases are Alport syndrome ($K_r = 1$ in Table 2) with 5 ingoing links and Kallmann syndrome ($K_r = 4$ in Table 2) with 3 ingoing links. Looking in more detail, Pfizer has the strongest links to Kallmann syndrome and Fabry disease. GlaxoSmithKline is linked to Fabry disease and medullary cystic kidney disease, while Bayer is oriented toward renal agenesis and familial renal amyloidosis disease. For Johnson & Johnson, the main orientations are toward Alport syndrome and Fabry disease. Novartis is mainly linked to renal tubular acidosis and Fabry disease. We observe that none of the articles devoted to the pharmaceutical companies with the largest market capitalization cite any of the rare renal diseases. All the connections from pharmaceutical companies to rare renal diseases are therefore indirect hidden links (red arrows in Fig 8), thereby indeed confirming the effective status of orphan diseases.

We also present the friendship network between companies and infectious diseases studied in [19] in SupInfo Fig S2.

Sensitivity of countries to pharmaceutical companies within Wikipedia

To see in Wikipedia, the influence of a specific pharmaceutical company on different countries we use the PageRank sensitivity D described in the Section Reduced Google matrix.

The PageRank sensitivity of an article devoted to country c to a specific pharmaceutical company article ph is determined by $D(ph \to c, c)$. This quantity measures the relative change of the country c article's PageRank probability caused by

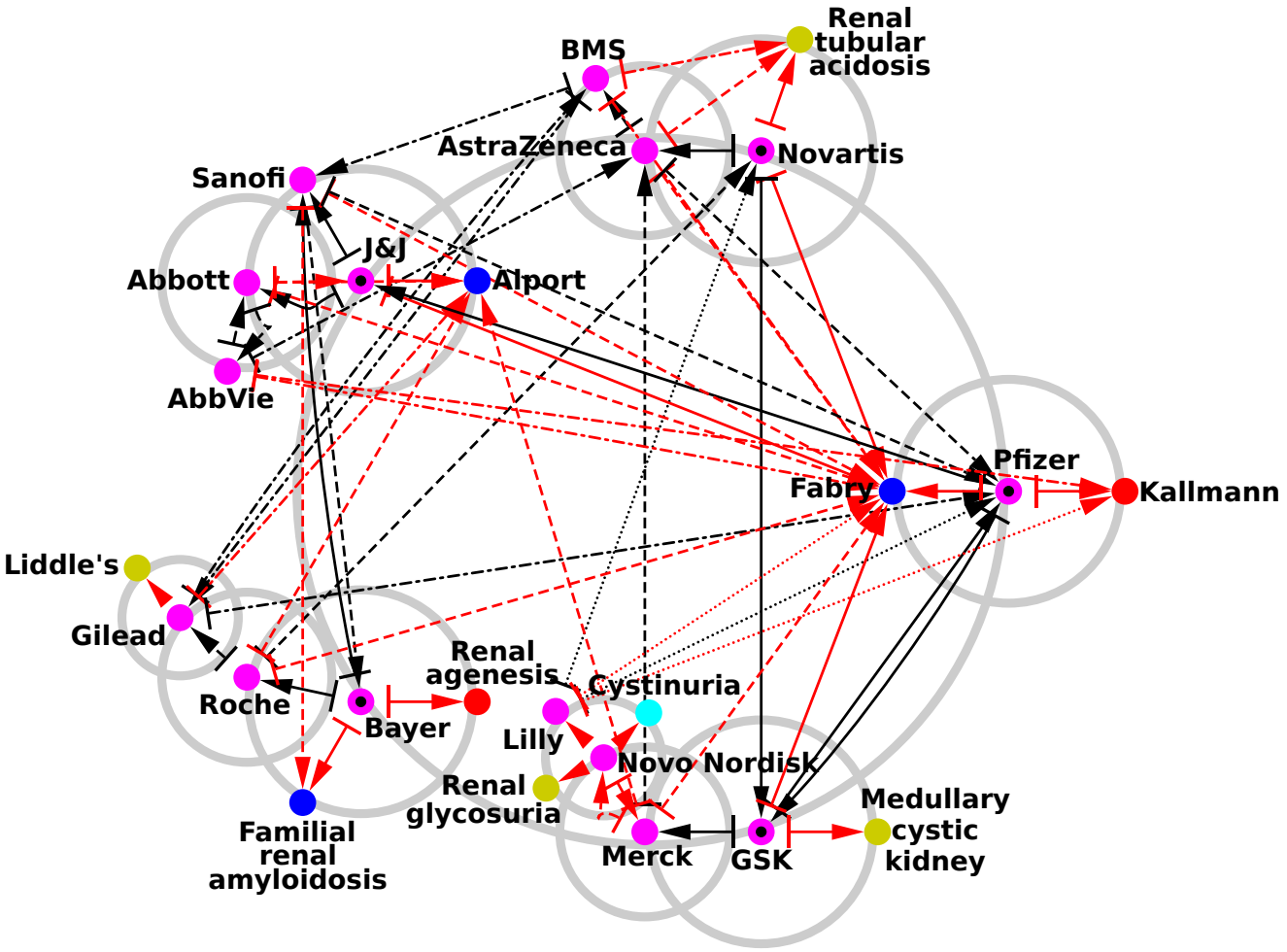

Fig 8. Reduced network of pharmaceutical companies with the addition of their best connected rare renal diseases. The construction algorithm is the same as the one used to generate Fig 6 excepting that we replace at each iteration the two best connected countries by the two best connected rare renal diseases. Pharmaceutical companies are represented by purple circles (●) and rare renal diseases by red circles (●) for congenital abnormalities of the kidney and urinary tract, blue circles (●) for glomerular diseases, gold circles (●) for renal tubular diseases and metabolic diseases, cyan circles (●) for nephrolithiasis, and green circles (●) for ciliopathies.

the increase in the intensity of the transition probability from the pharmaceutical company ph article to the country c article. In other words, the PageRank sensitivity $D(ph \to c, c)$ measures what happens to the influence of the country c article in Wikipedia when there is increased interest in the country c article resulting from the pharmaceutical company ph article. In the following, the expression *sensitivity of item A to item B* refers to the *PageRank sensitivity of the item A article to the item B article*. In Fig 9, we show the sensitivity of world countries to two companies, Pfizer and Bayer. The two most sensitive countries for Pfizer are Ireland and Italy. This observation correlates with the strongest direct links shown in Fig 6. Italy is present here for historical reasons as the calcium citrate used to produce citric acid was supplied to Pfizer by Italy until a shortage caused by World War I occurred, thus forcing Pfizer's chemists to develop fermentation technology to obtain citric acid from sugar by using a fungus. This technology was then used on a large scale to produce antibiotic penicillin

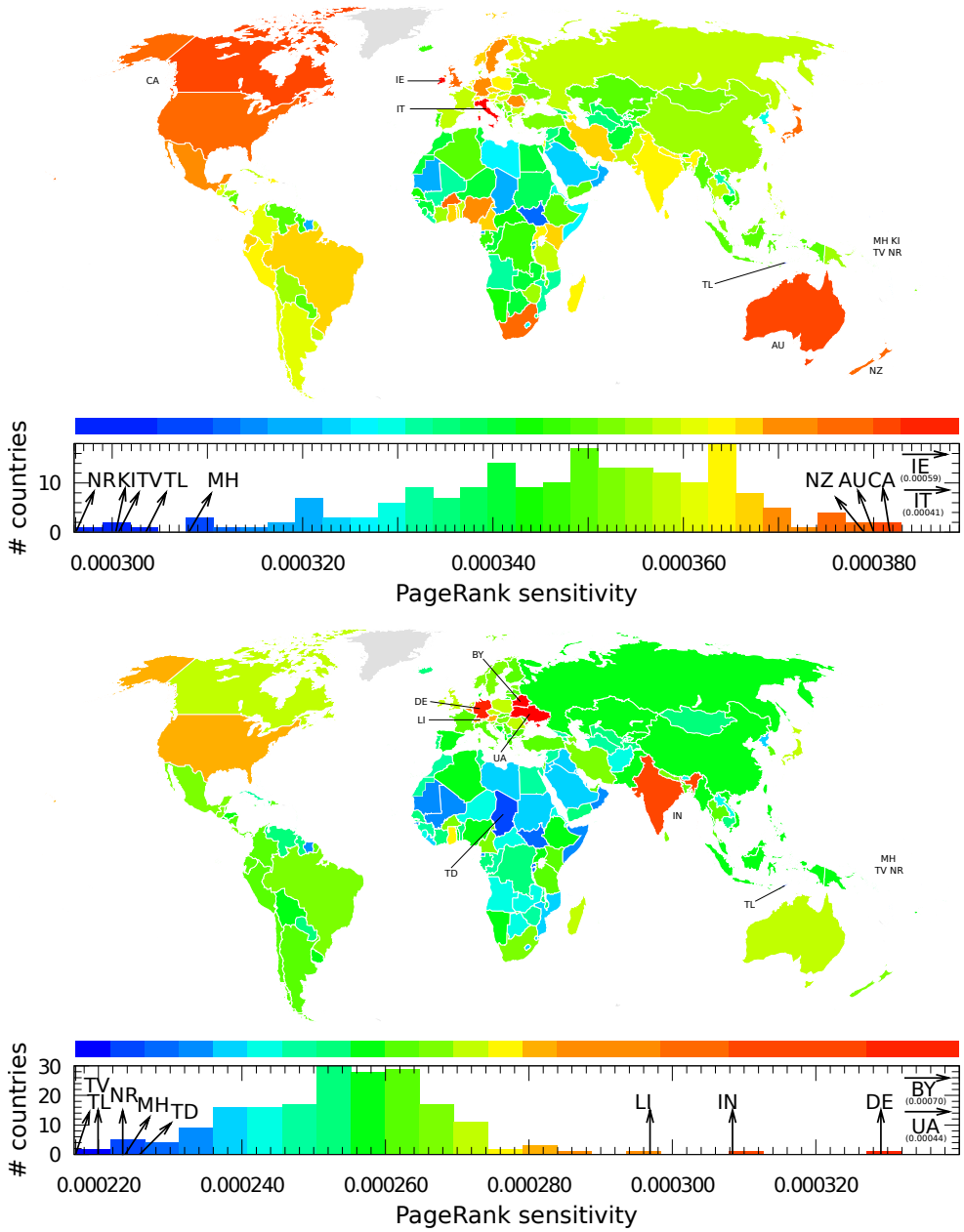

Fig 9. Sensitivity of countries to the Pfizer company (top panel) and the Bayer company (bottom panel). A country c is colored according to its diagonal PageRank sensitivity $D(ph \to c, c)$, where ph is the pharmaceutical company. Color categories are obtained using the Jenks natural breaks classification method [47].

during World War II [48]. In 2016, Pfizer attempted to acquire Allergan, an Irish–tax registered pharmaceutical company, in order to create the world's largest drugmaker and relocate its headquarters in Ireland to reduce taxes. The deal was called off after the US Treasury adopted new anti-tax inversion rules [49]. The next countries most impacted by Pfizer impacted countries are Canada, Australia, and New Zealand. There is no direct link between Pfizer and these countries. Among the shortest indirect paths, we have, e.g., Pfizer → Terre Haute → Canada; Pfizer → Wyeth → Australia; and

Pfizer → Helen Clark → Judicial Committee of the Privy Council → New Zealand. Pfizer has been present in Canada since the fifties, and in 2012, Pfizer's Canadian division was recognized as one of the 15 best places to work in Canada [50]. In 2009, Pfizer acquired Wyeth which supplied a pneumococcal vaccine approved for young children in Australia [51]. Wikileaks revealed that in 1990, Pfizer was lobbying in the US against New Zealand, whose drug-buying rules it considered to be restrictive; Helen Clark was at that time New Zealand Health Minister [52]. For Bayer the most sensitive countries are Belarus, Ukraine and Germany. Indeed, the article on Bayer has direct links to Belarus and Ukraine. There is also direct link to Germany and many indirect links to it since the company is located in Germany. It is possible that Germany is less sensitive to Bayer than Ukraine and Belarus are since its PageRank probability is significantly higher. We find again here Belarus and Ukraine, which are members of the CIS, as is explained the in Section Reduced network structures. India also appears as it hosts a part of Bayer Business Services [53].

Sensitivity of countries to rare renal diseases within Wikipedia

The PageRank sensitivity of a country c article to a specific rare renal disease rd is determined by $D(rd \to c, c)$. In Fig 10, we present the PageRank sensitivity of countries to Kallmann syndrome and Bardet–Biedl syndrome. For Kallmann syndrome, the most sensitive countries are Switzerland, Germany and Spain. From Kallmann syndrome, many paths converge toward Germany. We have, e.g., the direct link Kallmann syndrome → Germany, and, e.g., the indirect link Kallmann syndrome → Franz Josef Kallmann → Germany. Franz Josef Kallmann who gave the first description of this disease, was a German-born geneticist. The Wikipedia article devoted to Kallmann syndrome points to Switzerland through a direct link and an indirect link: Kallmann syndrome → Lausanne University Hospital → Switzerland. The Lausanne University Hospital is cited as one of the main research facilities working on this disease [54]. The Spanish doctor Aureliano Maestre de San Juan also identified the link between anosmia and hypogonadism (Kallmann syndrome → Aureliano Maestre de San Juan → Spain). Country sensitivity appears to be due to the combination of direct and indirect links captured by the REGOMAX algorithm.

In the case of Bardet–Biedl syndrome, the most sensitive countries are Burkina Faso, Norway and Romania. Here, there are no direct links to these countries from the article on Bardet–Biedl syndrome [55]. The indirect link to Romania appears because Arthur Biedl was born in what is now Romania. Indirect paths leading to the other most sensitive countries are difficult to find.

We also present the sensitivity of pharmaceutical companies to cancers and rare renal disease, respectively, in Fig S3 and Fig S4 in the Section Supplementary Information.

Discussion

We used the reduced Google matrix (REGOMAX) algorithm for an analysis of the English Wikipedia network with more than 5 million articles. The analysis focused on 195 world countries, the 34 largest biotechnology and pharmaceutical companies, 47 rare renal diseases and 37 types of cancer. The algorithm made it possible to construct the reduced Google matrix of these entries, while also taking account of the direct and indirect links between them. While the direct links were directly present in the global Wikipedia network, the indirect links were obtained with REGOMAX by summing the contributions between entries from all the pathways connecting them via the global network. Using the reduced Google matrix, we were able to determine the interaction

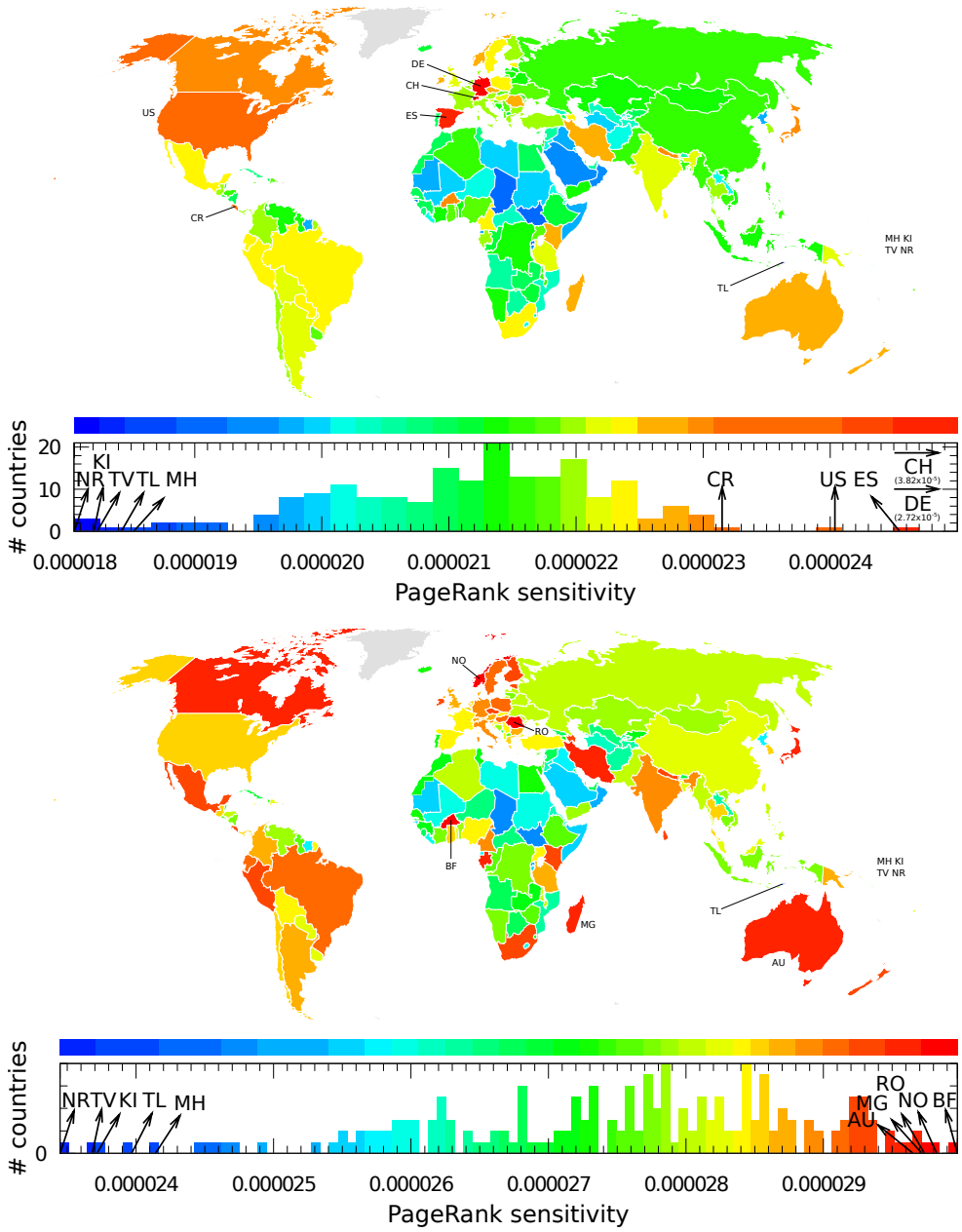

Fig 10. Sensitivity of countries to rare renal diseases; here to Kallmann syndrome (top panel) and to Bardet–Biedl syndrome (bottom panel). A country c is colored according to its diagonal PageRank sensitivity $D(rd \to c, c)$, where rd is the rare renal disease. Color categories are obtained using the Jenks natural breaks classification method [47].

networks between companies and countries, companies and rare renal diseases, and companies and cancers. The sensitivity of PageRank probabilities enabled us to identify the influence in Wikipedia of specific companies on countries. This approach also revealed the sensitivity of countries to specific rare renal diseases. We identified the most influential pharmaceutical companies and showed that the top PageRank positions are occupied by companies which are far from heading the market capitalization list.

We argue that improving Wikipedia articles about specific pharmaceutical companies can increase their worldwide visibility without significant additional investments. Our study shows that the knowledge accumulated in Wikipedia can be efficiently analyzed using the REGOMAX algorithm, which determines the effective interactions between specific Wikipedia articles that are of interest for researchers.

Acknowledgments

This work was partially supported by the Programme Investissements d'Avenir ANR-11-IDEX-0002-02, reference ANR-10-LABX-0037-NEXT (THETRACOM project). It was also partially supported the Programme Investissements d'Avenir ANR-15-IDEX-0003, ISITE-BFC (GNETWORKS project) and by the Bourgogne Franche-Comté region (APEX project).

Supplementary Information

In Fig S1 we present in the relative PageRank K_r – CheiRank K_r^* indexes plane the set of selected Wikipedia articles described in the Section Datasets. In Fig S2 we show the friendship network of interactions between pharmaceutical companies and infectious diseases studied in [19].

We also show that the REGOMAX analysis allows us to determine the inter-sensitivity of pharmaceutical companies to cancers and rare renal diseases. We show the sensitivity of 34 pharmaceutical companies to 37 cancers and 47 rare renal diseases (and vice versa) in Fig S3 and Fig S4 respectively. We note that the very strong matrix element in Fig S4 (red square) between Shire and Fabry disease appears due to the drug Replagal which is a treatment for this disease and is produced by Shire.

References

1. World Health Organization. World Health Statistics; 2018. Available from: https://www.who.int/gho/publications/world_health_statistics/.

2. Wikipedia contributors. List of largest biotechnology and pharmaceutical companies — Wikipedia, The Free Encyclopedia; 2019. Available from: https://en.wikipedia.org/w/index.php?title=List_of_largest_biotechnology_and_pharmaceutical_companies&oldid=884151571.

3. Giles J. Internet encyclopaedias go head to head. Nature. 2005;438:900–901. doi:10.1038/438900a.

4. Callaway E. No rest for the bio-wikis. Nature. 2010;468(7322):359–360. doi:10.1038/468359a.

5. Nielsen FÅ. Wikipedia Research and Tools: Review and Comments. SSRN Electronic Journal. 2012;doi:10.2139/ssrn.2129874.

6. Lewoniewski W, Wecel K, Abramowicz W. Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. Informatics. 2017;4(4):43. doi:10.3390/informatics4040043.

7. Dorogovtsev S. Lectures on Complex Networks. Oxford University Press; 2010.

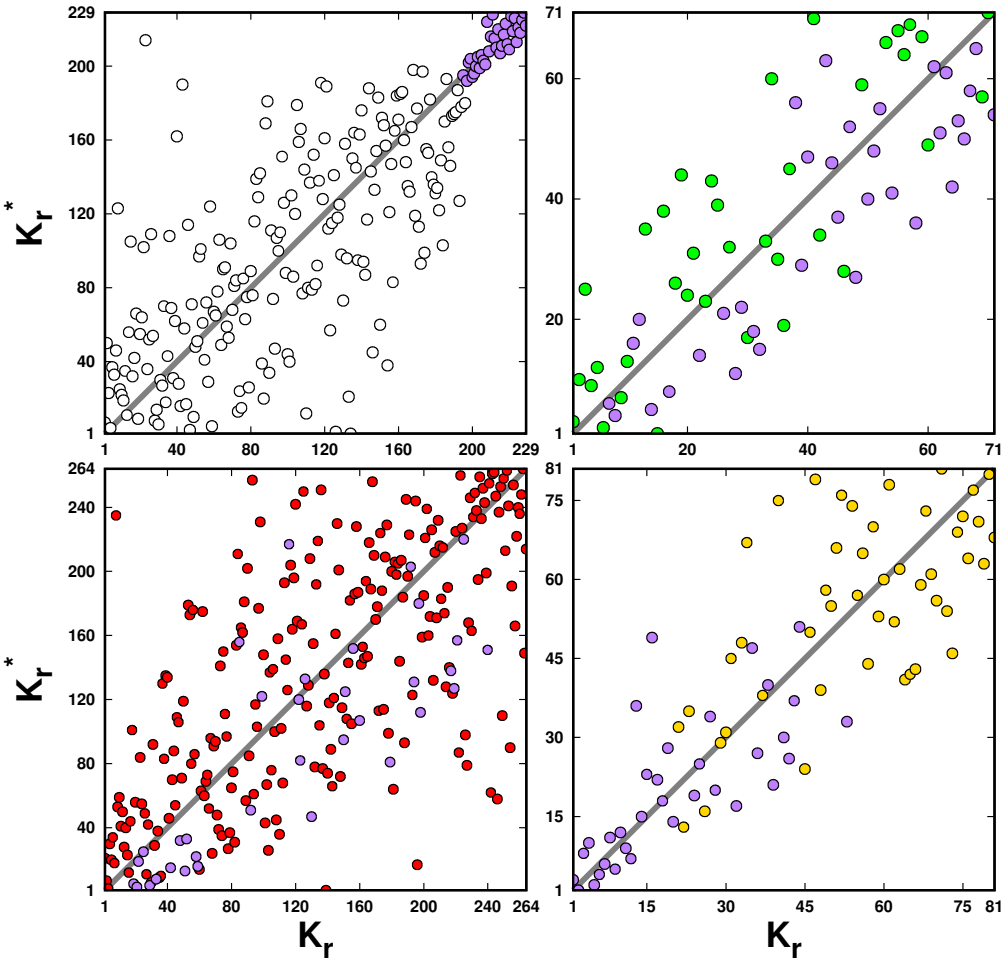

Fig S1. Distribution of the May 2017 English Wikipedia articles in the relative PageRank K_r – CheiRank K_r^* indexes plane for pharmaceutical companies and countries (top left panel), cancer types (top right panel), infectious diseases (bottom left panel), and rare renal diseases (bottom right panel). The $N_{ph} = 34$ pharmaceutical companies are represented by purple circles (●), the $N_c = 195$ countries by white circles (◯), the $N_{cr} = 37$ cancer types by green circles (●), the $N_d = 230$ infectious diseases by red circles (●), and the $N_{rd} = 47$ rare renal diseases by gold circles (●).

8. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems. 1998;30(1):107 – 117. doi:10.1016/S0169-7552(98)00110-X.

9. Langville AN, Meyer CD. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press; 2012.

10. Ermann L, Frahm KM, Shepelyansky DL. Google matrix analysis of directed networks. Rev Mod Phys. 2015;87:1261–1310. doi:10.1103/RevModPhys.87.1261.

11. Eom YH, Aragón P, Laniado D, Kaltenbrunner A, Vigna S, Shepelyansky DL. Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. PLOS ONE. 2015;10(3):1–27. doi:10.1371/journal.pone.0114825.

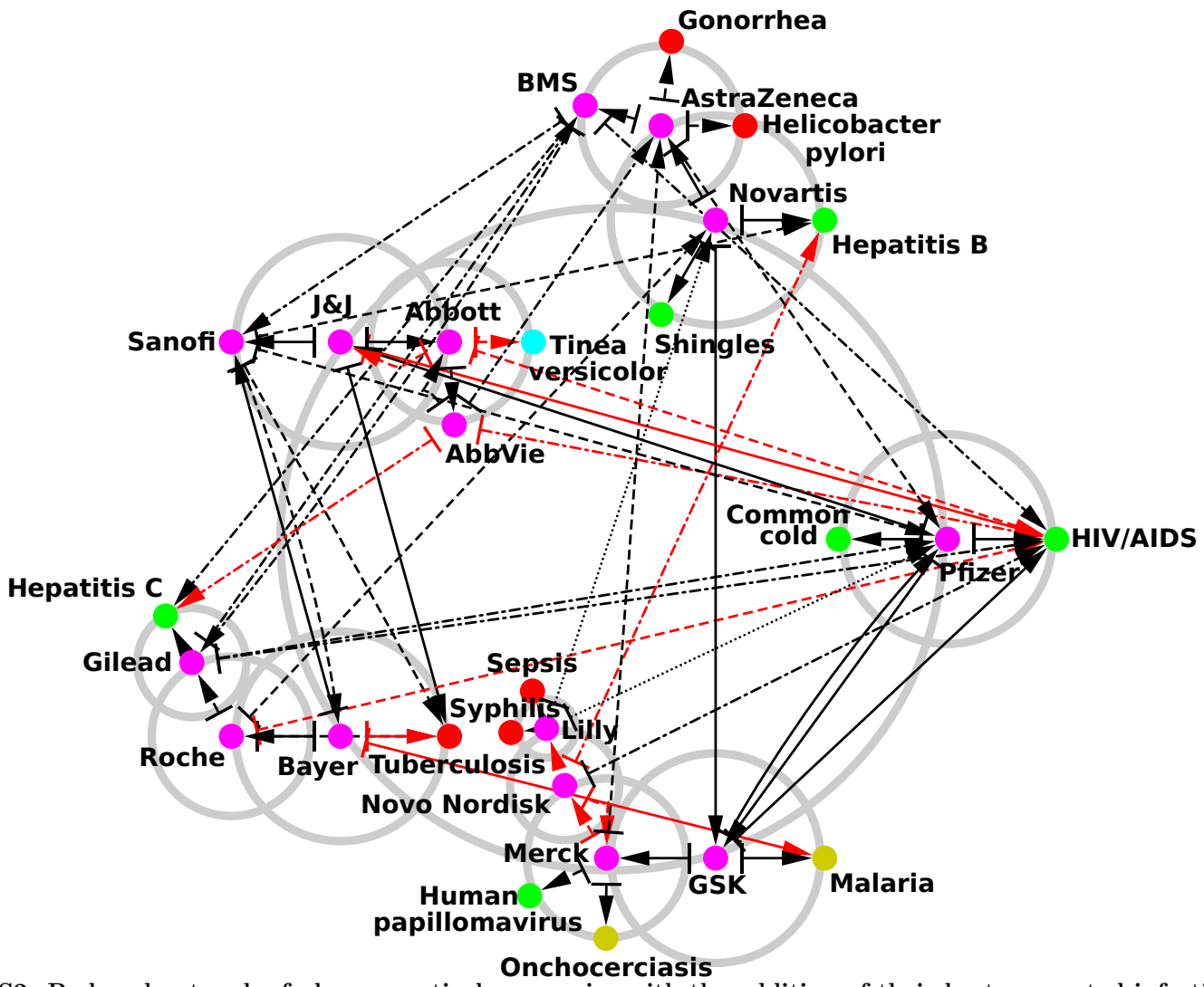

Fig S2. Reduced network of pharmaceutical companies with the addition of their best connected infectious diseases. The construction algorithm is the same as the one used to generate Fig 6 excepting that we replace at each iteration the two best connected countries by the two best connected infectious diseases. Pharmaceutical companies are represented by purple circles (●) and infectious diseases by red circles (●) for bacterial type diseases, green circles (●) for viral type diseases, gold circles (●) for parasitic type diseases, and cyan circles (●) for fungal type diseases. The list of infectious diseases is available in [19].

12. Lages J, Patt A, Shepelyansky DL. Wikipedia ranking of world universities. The European Physical Journal B. 2016;89(3). doi:10.1140/epjb/e2016-60922-0.

13. Frahm KM, Shepelyansky DL. Reduced Google matrix. arXiv. 2016;arXiv:1602.02394.

14. Frahm KM, Jaffrès-Runser K, Shepelyansky DL. Wikipedia mining of hidden links between political leaders. The European Physical Journal B. 2016;89(12):269. doi:10.1140/epjb/e2016-70526-3.

15. El Zant S, Jaffrès-Runser K, Shepelyansky DL. Capturing the influence of geopolitical ties from Wikipedia with reduced Google matrix. PLOS ONE.

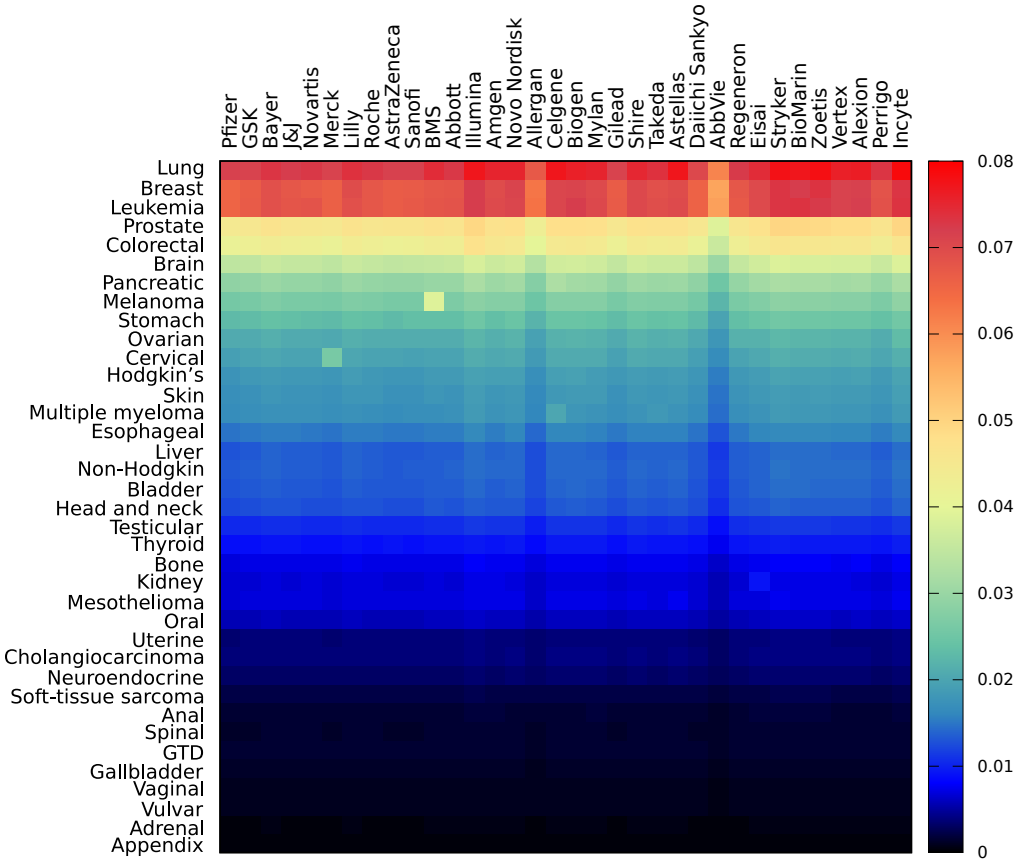

Fig S3. Sensitivity of pharmaceutical companies to cancers. Horizontal (vertical) entries represents pharmaceutical companies (cancer types). The acronym GTD stands for gestational trophoblastic disease.

2018;13(8):1–31. doi:10.1371/journal.pone.0201397.

16. Coquidé C, Lages J, Shepelyansky DL. World influence and interactions of universities from Wikipedia networks. The European Physical Journal B. 2019;92(1):3. doi:10.1140/epjb/e2018-90532-7.

17. Demidov D, Frahm KM, Shepelyansky DL. What is the central bank of Wikipedia? arXiv e-prints. 2019; p. arXiv:1902.07920.

18. Lages J, Shepelyansky DL, Zinovyev A. Inferring hidden causal relations between pathway members using reduced Google matrix of directed biological networks. PLOS ONE. 2018;13(1):1–28. doi:10.1371/journal.pone.0190812.

19. Rollin G, Lages J, Shepelyansky DL. World Influence of Infectious Diseases From Wikipedia Network Analysis. IEEE Access. 2019;7:26073–26087. doi:10.1109/ACCESS.2019.2899339.

20. Rollin G, Lages J, Shepelyansky D. Wikipedia network analysis of cancer interactions and world influence. bioRxiv. 2019;doi:10.1101/527879.

21. International Agency for Research on Cancer, World Health Organization. GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012 v1.0; 2012. Available from:

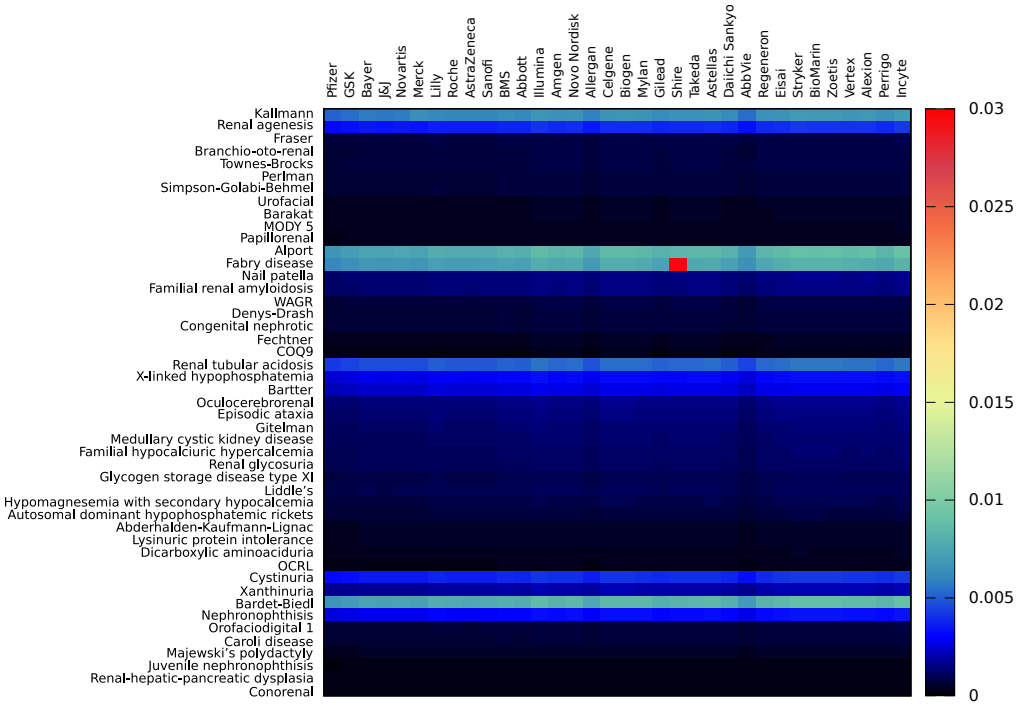

Fig S4. Sensitivity of pharmaceutical companies to rare renal diseases. Horizontal (vertical) entries represents pharmaceutical companies (rare renal diseases).

http://publications.iarc.fr/Databases/Iarc-Cancerbases/GLOBOCAN-2012-Estimated-Cancer-Incidence-Mortality-And-Prevalence-Worldwi 0-2012.

22. Global Genes. RARE Facts; 2019. Available from: https://globalgenes.org/rare-facts/.

23. Kaplan W, Wirtz VJ, Mantel-Teeuwisse A, Stolk P, Duthey B, Laing R. Priority Medicines for Europe and the World 2013 Update; 2013. Available from: https://www.who.int/medicines/areas/priority_medicines/Ch6_19Rare.pdf.

24. National Cancer Institute. Cancer types; Accessed March 2019. Available from: https://www.cancer.gov/types.

25. Orphanet. The portal for rare diseases and orphan drugs; Accessed: March 2019. https://www.orpha.net.

26. OMIM. Online Mendelian Inheritance in Man: An Online Catalog of Human Genes and Genetic Disorders; Accessed: March 2019. https://www.omim.org.

27. Devuyst O, Knoers NVAM, Remuzzi G, Schaefer F. Rare inherited kidney diseases: challenges, opportunities, and perspectives. The Lancet. 2014;383:1844–1859. doi:10.1016/S0140-6736(14)60659-0.

28. Chepelianskii AD. Towards physical laws for software architecture. arXiv e-prints. 2010; p. arXiv:1003.5455.

29. Zhirov AO, Zhirov OV, Shepelyansky DL. Two-dimensional ranking of Wikipedia articles. The European Physical Journal B. 2010;77(4):523–531. doi:10.1140/epjb/e2010-10500-7.

30. Rollin G, Lages J, Serebriyskaya TS, Shepelyansky D. WikiPharma: Wikipedia network of pharmaceutical companies; 2019. Available from: http://perso.utinam.cnrs.fr/~lages/datasets/WikiPharma/.

31. Alexa. World top visited web sites; 2019. Available from: https://www.alexa.com/topsites.

32. Scolari F, Ghiggeri G. Nephronophthisis-Medullary Cystic Kidney Disease: From Bedside to Bench and Back Again. Saudi Journal of Kidney Diseases and Transplantation. 2003;14(3):316–327.

33. Tsuboi K, Yamamoto H. Clinical observation of patients with Fabry disease after switching from agalsidase beta (Fabrazyme) to agalsidase alfa (Replagal). Genetics In Medicine. 2012;14:779. doi:10.1038/gim.2012.39.

34. Wikipedia contributors. Ludwig N. Hantson — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Ludwig_N._Hantson&oldid=773365464.

35. Bloomberg. Ludwig N. Hantson: Executive Profile and Biography; 2019. Available from: https://www.bloomberg.com/research/stocks/people/person.asp?personId=12703772&privcapId=347983.

36. Wikipedia contributors. Baxalta — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Baxalta&oldid=768715356.

37. Wikipedia contributors. Abbott Medical Optics — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Abbott_Medical_Optics&oldid=767847481.

38. Wikipedia contributors. Hypocalciuria — Wikipedia, The Free Encyclopedia; 2016. Available from: https://en.wikipedia.org/w/index.php?title=Hypocalciuria&oldid=749626627.

39. Wikipedia contributors. Sensorineural hearing loss — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Sensorineural_hearing_loss&oldid=770911856.

40. Wikipedia contributors. Cataract — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Cataract&oldid=777851128.

41. Astuti D, Morris MR, Cooper WN, Staals RHJ, Wake NC, Fews GA, et al. Germline mutations in DIS3L2 cause the Perlman syndrome of overgrowth and Wilms tumor susceptibility. Nature Genetics. 2012;44:277–284. doi:10.1038/ng.1071.

42. Drash A, Sherman F, Hartmann WH, Blizzard RM. A syndrome of pseudohermaphroditism, Wilms' tumor, hypertension, and degenerative renal disease. The Journal of Pediatrics. 1970;76(4):585 – 593. doi:https://doi.org/10.1016/S0022-3476(70)80409-7.

43. Alexion. Alexion to Acquire Synageva to Strengthen Global Leadership in Developing and Commercializing Transformative Therapies for Patients with Devastating and Rare Diseases; 2019. Available from: https://news.alexion.com/node/524/all/2006/all.

44. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003;13(11):2498–2504. doi:10.1101/gr.1239303.

45. Bayer. Bayer HealthCare successfully completes acquisition of Sagmel, Inc.'s OTC Business; 2019. Available from: https://www.investor.bayer.com/en/nc/news/archive/investor-news-2008/investor-news-2008/bayer-healthcare-successfully-completes-acquisition-of-sagmel-incs-otc-bu

46. Bayer. Bayer HealthCare to acquire OTC Business of Sagmel, Inc.; 2019. Available from: https://www.investor.bayer.de/en/nc/news/archive/investor-news-2008/investor-news-2008/bayer-healthcare-to-acquire-otc-business-of-sagmel-inc/.

47. Wikipedia contributors. Jenks natural breaks optimization — Wikipedia, The Free Encyclopedia; 2018. Available from: https://en.wikipedia.org/w/index.php?title=Jenks_natural_breaks_optimization&oldid=842502988.

48. ACS Chemistry for Life. Pfizer's work on penicillin for World War II becomes a National Historic Chemical Landmark; 2019. Available from: https://www.acs.org/content/acs/en/pressroom/newsreleases/2008/june/pfizers-work-on-penicillin-for-world-war-ii-becomes-a-national-historic-c html.

49. The Washington Post. Pfizer, Allergan call off $160 billion merger after U.S. moves to block inversions; 2019. Available from: http://wapo.st/1VahSX7?tid=ss_mail&utm_term=.2d3472c4dcf9.

50. Wikipedia contributors. Pfizer — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Pfizer&oldid=777843272.

51. Wikipedia contributors. Wyeth — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Wyeth&oldid=777064151.

52. New Zealand Herald. WikiLeaks: Drug firms tried to ditch Clark; 2019. Available from: https://www.nzherald.co.nz/pharmaceuticals/news/article.cfm?c_id=278&objectid=10695239.

53. Wikipedia contributors. Bayer — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Bayer&oldid=776044035.

54. Wikipedia contributors. Kallmann syndrome — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Kallmann_syndrome&oldid=776085509.

55. Wikipedia contributors. Bardet–Biedl syndrome — Wikipedia, The Free Encyclopedia; 2017. Available from: https://en.wikipedia.org/w/index.php?title=Bardet%E2%80%93Biedl_syndrome&oldid=775800611.