Open Archive Toulouse Archive Ouverte

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible

Official URL:

https://doi.org/10.1016/j.physleta.2017.06.021

Multi-cultural Wikipedia mining of geopolitics interactions leveraging reduced Google matrix analysis

Klaus M. Frahm [a], Samer El Zant [b], Katia Jaffrès-Runser [b], Dima L. Shepelyansky [a],*

[a] *Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, F-31062 Toulouse, France*
[b] *Institut de Recherche en Informatique de Toulouse, Université de Toulouse, INPT, 31061 Toulouse, France*

ABSTRACT

Geopolitics focuses on political power in relation to geographic space. Interactions among world countries have been widely studied at various scales, observing economic exchanges, world history or international politics among others. This work exhibits the potential of Wikipedia mining for such studies. Indeed, Wikipedia stores valuable fine-grained dependencies among countries by linking webpages together for diverse types of interactions (not only related to economical, political or historical facts). We mine herein the Wikipedia networks of several language editions using the recently proposed method of reduced Google matrix analysis. This approach allows to establish direct and hidden links between a subset of nodes that belong to a much larger directed network. Our study concentrates on 40 major countries chosen worldwide. Our aim is to offer a multicultural perspective on their interactions by comparing networks extracted from five different Wikipedia language editions, emphasizing English, Russian and Arabic ones. We demonstrate that this approach allows to recover meaningful direct and hidden links among the 40 countries of interest.

1. Introduction

Political and economic interactions between regions of the world have always been of utmost interest to measure and predict their relative influence. Such studies belong to the field of geopolitics that focuses on political power in relation to geographic space. Interactions among world countries have been widely studied at various scales (worldwide, continental or regional) using different types of information. Studies are driven by observing economic exchanges, social changes, history, international politics and diplomacy among others [1,2]. The major finding of this paper is to show that meaningful worldwide interactions can be automatically extracted from the global and free online Encyclopaedia Wikipedia [3] for a given set of countries. All information gathered in this collaborative knowledge base can be leveraged to provide a picture of countries relationships, fostering a new framework for thorough geopolitics studies.

Wikipedia has become the largest open source of knowledge being close to Encyclopaedia Britannica [4] by the accuracy of its scientific entries [5] and overcoming the latter by the enormous quantity of available information. A detailed analysis of strong and weak features of Wikipedia is given at [6,7]. Wikipedia articles make citations to each other, providing a direct relationship between webpages and topics. As such, Wikipedia generates a larger directed network of article titles with a rather clear meaning. For these reasons, it is interesting to apply algorithms developed for search engines of World Wide Web (WWW) such as the PageRank algorithm [8] (see also [9]), to analyze the ranking properties and relations between Wikipedia articles. For various language editions of Wikipedia it was shown that the PageRank vector produces a reliable ranking of historical figures over 35 centuries of human history [10–14] and a solid Wikipedia ranking of world universities (WRWU) [10,15]. It has been shown that the Wikipedia ranking of historical figures is in a good agreement with the well-known Hart ranking [16], while the WRWU is in a good agreement with the Shanghai Academic ranking of world universities [17].

At present directed networks of real systems can be very large (about 4.2 million articles for the English Wikipedia edition in 2013 [13] or 3.5 billion web pages (called also nodes) for a publicly accessible web crawl that was gathered by the Common Crawl Foundation in 2012 [18]). For some studies, one might be interested only in the particular interactions between a very small subset of nodes compared to the full network size. For instance, in

* Corresponding author.
  E-mail addresses: frahm@irsamc.ups-tlse.fr (K.M. Frahm),
samer.elzant@enseeiht.fr (S. El Zant), kjr@enseeiht.fr (K. Jaffrès-Runser),
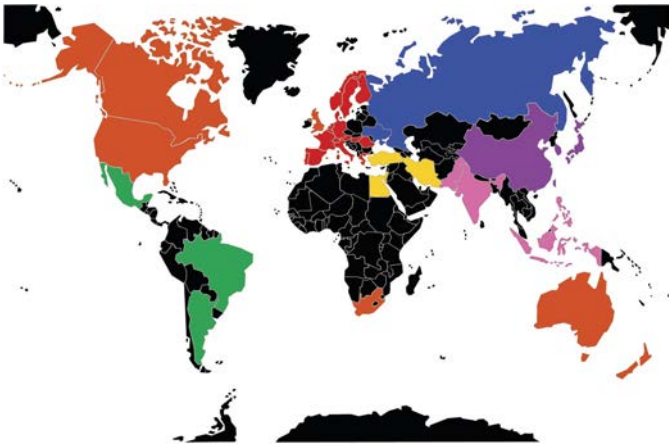dima@irsamc.ups-tlse.fr (D.L. Shepelyansky).

Fig. 1. Geographical distribution of the 40 selected countries. Color code groups countries into 7 sets: orange (OC) for English speaking countries, blue (BC) for former Soviet union ones, red (RC) for European ones, green (GC) for Latin American ones, yellow (YC) for Middle Eastern ones, purple (PUC) for North-East Asian ones and finally pink (PIC) for South-Eastern countries (see colors and country names in Table 1; other countries are shown in black). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
List of names of 40 selected countries with PageRank K, CheiRank K^* for EnWiki, ArWiki and RuWiki, ordered by increasing PageRank of EnWiki edition. Fig. 1 gives color correspondence details. A Wikipedia article with country name represents one node of the whole network with N nodes, e.g. https://en.wikipedia.org/wiki/France for "France" in EnWiki. (For interpretation of the references to color in this table, the reader is referred to the web version of this article.)

Country	EnWiki		ArWiki		RuWiki	
	K	K^*	K	K^*	K	K^*
United States (US) OC	1	9	1	5	2	27
France (FR) RC	2	19	3	31	3	14
United Kingdom (UK) OC	3	25	6	20	7	3
Germany (DE) RC	4	33	8	14	4	24
Canada (CA) OC	5	26	13	19	12	26
India (IN) PIC	6	23	9	25	13	8
Australia (AU) OC	7	35	16	22	18	12
Italy (IT) RC	8	15	5	1	6	32
Japan (JP) PUC	9	4	11	9	11	7
China (CN) PUC	10	8	12	17	9	21
Russia (RU) BC	11	6	7	2	1	2
Spain (ES) RC	12	30	4	8	8	15
Poland (PL) RC	13	12	26	32	10	17
The Netherlands (NL) RC	14	37	18	33	15	31
Iran (IR) YC	15	2	14	15	30	22
Brazil (BR) GC	16	3	21	26	20	1
Sweden (SE) RC	17	22	22	7	19	5
New Zealand (NZ) OC	18	28	34	24	36	4
Mexico (MX) GC	19	40	23	38	22	37
Switzerland (CH) RC	20	38	20	34	16	18
Norway (NO) RC	21	32	35	16	27	11
Romania (RO) RC	22	10	19	6	32	36
Turkey (TR) YC	23	7	15	13	21	38
South Africa (ZA) OC	24	24	29	39	35	20
Belgium (BE) RC	25	18	27	37	29	30
Austria (AT) RC	26	39	28	28	14	28
Greece (GR) RC	27	21	10	36	25	25
Argentina (AR) GC	28	1	32	29	33	23
Philippines (PH) PIC	29	17	36	21	39	33
Portugal (PT) RC	30	36	24	12	17	9
Pakistan (PK) PUC	31	5	25	35	37	29
Denmark (DK) RC	32	16	33	10	31	19
Israel (IL) YC	33	20	17	18	28	6
Finland (FI) RC	34	14	37	4	26	16
Egypt (EG) YC	35	31	2	3	24	39
Indonesia (ID) PIC	36	13	31	11	34	10
Hungary (HU) RC	37	11	40	40	23	40
Taiwan (TW) PUC	38	27	39	27	40	34
South Korea (KR) PUC	39	34	38	30	38	35
Ukraine (UA) BC	40	29	30	23	5	13

this paper, we are interested in capturing the interactions of the 40 countries represented in Fig. 1 using the networks extracted from five Wikipedia language editions covering a few millions of articles each. However, let us assume that there is a rather important person (having his own Wikipedia article corresponding to a node C) who was born in country A and worked the main part of his life in country B; therefore A and B may have links to C (in either direction) and thus there may be an indirect link between the two nodes A and B via the node C (or other nodes). In previous works, a solution to this general problem has been proposed in [19,20] by defining the reduced Google matrix theory. Main elements of Reduced Google matrix G_R will be presented next, but in a few words, it captures in a 40-by-40 Perron–Frobenius matrix the full contribution of direct and indirect interactions happening in the full Google matrix between the 40 nodes of interest (we took top 40 countries of PageRank vector of EnWiki). Elements of reduced matrix $G_R(i, j)$ can be interpreted as the probability for a random surfer starting at webpage j to arrive in webpage i using direct and indirect interactions. Indirect interactions refer to paths composed in part of webpages different from the 40 ones of interest. Even more interesting and unique to reduced Google matrix theory, we show here that intermediate computation steps of G_R offer a decomposition of G_R into matrices that clearly distinguish direct from indirect interactions. As such, it is possible to extract the probability for an indirect interaction between two nodes to happen.

Reduced Google matrix theory is a perfect candidate for analyzing the direct and indirect interactions between countries selected worldwide. In this paper, we extract from G_R and its decomposition into direct and indirect matrices of selected subset network of $N_r = 40$ countries. The Google matrix of this subset network of N_r nodes is computed taking into account direct and hidden (i.e. indirect) directed links. More specifically, we deduce a fine-grained classification of countries that captures what we call the *hidden friends* and *hidden followers* of a given country. The structure of these graphs provides relevant social information: communities of countries with strong ties can be clearly exhibited while countries acting as bridges are present as well. This is mainly the case for the hidden interactions networks of friends (or followers) that offer new information compared to the direct networks of friends (or followers) whose topology is mainly enforced by top PageRank countries. The mathematical procedure of the reduced

G_R matrix construction for N_r nodes is described in detail in Section 2.

The networks of G_R direct and hidden interactions can be calculated for different Wikipedia language editions. In this paper, reduced Google matrix analysis is applied to the same set of 40 countries on networks representing five different Wikipedia editions: English (EnWiki), Arabic (ArWiki), Russian (RuWiki), French (FrWiki) and German (DeWiki) editions. We take for analysis the top 40 countries according to the EnWiki PageRank. Wikipedia language editions are usually modified by authors who mainly belong to the region associated with this language. Thus our study shows the impact of this cultural bias when comparing direct and hidden networks of friends (or followers) among different language editions. We show that part of the interactions are cross-cultural while others are clearly biased by the culture of the authors.

In Section 2 we introduce the main elements of reduced Google matrix theory, Section 3 describes G_R calculated for 40 countries and for five different Wikipedia editions. Specific emphasis is given to the very different English, Arabic and Russian editions. Networks of friends and followers for direct and hidden interaction matrices are created and discussed in Section 4, and conclusion is drawn in Section 5.
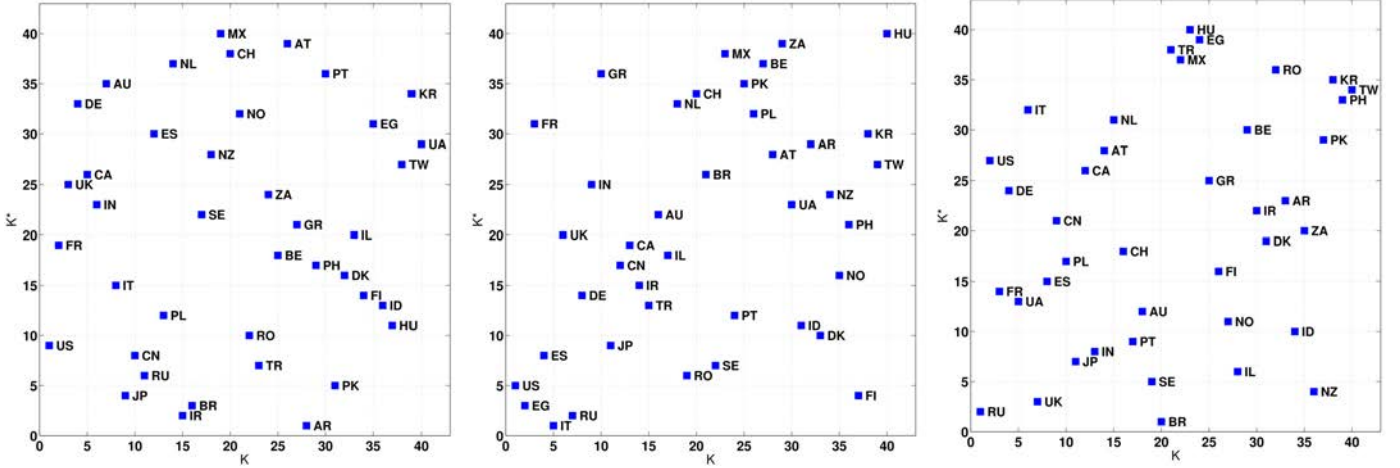
Fig. 2. Position of countries in the local (K, K^*) plane of the reduced network of 40 countries in the EnWiki (left), ArWiki (middle) and RuWiki (right) networks.

2. Reduced Google matrix theory

2.1. Google matrix

It is convenient to describe the network of N Wikipedia articles: a network node is given by the article name, all nodes are numbered, some nodes correspond to articles with country names, see Table 1; the number of countries is much smaller than the total number of articles N. Then the Google matrix G is constructed from the adjacency matrix A_{ij} with elements 1 if article (node) j points to article (node) i and zero otherwise. In this case, elements of the Google matrix take the standard form [8,9]

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

where S is the matrix of Markov transitions with elements $S_{ij} = A_{ij}/k_{out}(j)$, $k_{out}(j) = \sum_{i=1}^N A_{ij} \neq 0$ being the node j out-degree (number of outgoing links) and with $S_{ij} = 1/N$ if j has no outgoing links (dangling node). Here $0 < \alpha < 1$ is the damping factor which for a random surfer determines the probability $(1-\alpha)$ to jump to any node; below we use $\alpha = 0.85$. The right eigenvectors $\psi_i(j)$ of G are defined by:

$$\sum_{j'} G_{jj'} \psi_i(j') = \lambda_i \psi_i(j) \ . \tag{2}$$

The PageRank eigenvector $P(j) = \psi_{i=0}(j)$ corresponds to the largest eigenvalue $\lambda_{i=0} = 1$ [8,9]. It has positive elements which give the probability to find a random surfer on a given node in the stationary long time limit of the Markov process. All nodes can be ordered by a monotonically decreasing probability $P(K_i)$ with the highest probability at $K = 1$. The index K is the PageRank index. The PageRank vector is computed by the PageRank algorithm with iterative multiplication of initial random vector by G matrix [8,9,14]. Left eigenvectors are biorthogonal to right eigenvectors of different eigenvalues. The left eigenvector for $\lambda = 1$ has identical (unit) entries due to the column sum normalization of G. In the following we use the notations ψ_L^T and ψ_R for left and right eigenvectors, respectively. Notation T stands for vector or matrix transposition.

In addition to the matrix G it is useful to introduce a Google matrix G^* constructed from the adjacency matrix of the same network but with inverted direction of all links [21]. The vector $P^*(K^*)$ is called the CheiRank vector [10,21] and the index numbering nodes in order of monotonic decrease of probability P^* is noted as CheiRank index K^*. Thus, nodes with many ingoing (or outgoing) links have small values of $K = 1, 2, 3\ldots$ (or of

$K^* = 1, 2, 3, \ldots$) [9,14]. We show the distribution of selected countries on the PageRank–ChiRank plane (K, K^*) in Fig. 2.

2.2. Reduced Google matrix

We construct the reduced Google matrix for a certain subset of N_r selected nodes from the global Wikipedia network with N nodes ($N_r \ll N$). As a subset we choose top 40 countries with the largest PageRank probabilities for EnWiki network (the names are given in Table 1). The reduced Google matrix G_R is constructed on the mathematical basis described below. The main element of this construction is to keep the same PageRank probabilities of N_r nodes as in the global network (up to a fixed multiplier coefficient) and to take into account all indirect links between N_r nodes coupled by transitions via $N - N_r$ nodes of the global network (see also [20]).

Let G be a typical Google matrix (1) for a network with N nodes such that $G_{ij} \geq 0$ and the column sum normalization $\sum_{i=1}^N G_{ij} = 1$ is verified. We consider a sub-network with $N_r < N$ nodes, called "reduced network". In this case we can write G in a block form:

$$G = \begin{pmatrix} G_{rr} & G_{rs} \\ G_{sr} & G_{ss} \end{pmatrix} \tag{3}$$

where the index "r" refers to the nodes of the reduced network and "s" to the other $N_s = N - N_r$ nodes which form a complementary network which we will call the "scattering network". Thus G_{rr} is given by the direct links between selected N_r nodes, G_{ss} describes links between all other $N - N_r$ nodes, G_{rs} and G_{sr} give links between these two parts. PageRank vector of the full network is given by:

$$P = \begin{pmatrix} P_r \\ P_s \end{pmatrix} \tag{4}$$

which satisfies the equation $GP = P$ or in other words P is the right eigenvector of G for the unit eigenvalue. This eigenvalue equation reads in block notations:

$$(1 - G_{rr}) P_r - G_{rs} P_s = 0, \tag{5}$$

$$-G_{sr} P_r + (1 - G_{ss}) P_s = 0. \tag{6}$$

Here **1** is the unit matrix of corresponding size N_r or N_s. Assuming that the matrix $1 - G_{ss}$ is not singular, i.e. all eigenvalues G_{ss} are strictly smaller than unity (in modulus), we obtain from (6) that

$$P_s = (1 - G_{ss})^{-1} G_{sr} P_r \tag{7}$$

which gives together with (5):

$$G_R P_r = P_r, \qquad G_R = G_{rr} + G_{rs}(1 - G_{ss})^{-1}G_{sr} \qquad (8)$$

where the matrix G_R of size $N_r \times N_r$, defined for the reduced network, can be viewed as an effective reduced Google matrix. Here the contribution of G_{rr} accounts for direct links in the reduced network and the second matrix inverse term corresponds to all contributions of indirect links of arbitrary order. The matrix elements of G_R are non-negative since the matrix inverse in (8) can be expanded as:

$$(1 - G_{ss})^{-1} = \sum_{l=0}^{\infty} G_{ss}^l \quad . \qquad (9)$$

In (9) the integer l represents the order of indirect links, i.e. the number of indirect links which are used to connect indirectly two nodes of the reduced network. We refer the reader to [20] to get the proof that G_R also fulfills the condition of column sum normalization being unity.

2.3. Numerical evaluation of G_R

We can question how to evaluate practically the expression (8) of G_R for a particular sparse and quite large network when $N_r \sim 10^2$–10^3 is small compared to N and $N_s \approx N \gg N_r$. If N_s is too large (e.g. $N_s > 10^5$) a direct naive evaluation of the matrix inverse $(1 - G_{ss})^{-1}$ in (8) by Gauss algorithm is not efficient. In this case we can try the expansion (9) provided it converges sufficiently fast with a modest number of terms. However, this is most likely not the case for typical applications since G_{ss} is very likely to have at least one eigenvalue very close to unity.

Therefore, we consider the situation where the full Google matrix has a well defined gap between the leading unit eigenvalue and the second largest eigenvalue (in modulus). For example if G is defined using a damping factor α in the standard way, as in (1), the gap is at least $1 - \alpha$ which is 0.15 for the standard choice $\alpha = 0.85$ [9]. In order to evaluate the expansion (9) efficiently, we need to take out analytically the contribution of the leading eigenvalue of G_{ss} close to unity which is responsible for the slow convergence.

Below we denote by λ_c this leading eigenvalue of G_{ss} and by ψ_R (ψ_L^T) the corresponding right (left) eigenvector such that $G_{ss}\psi_R = \lambda_c \psi_R$ (or $\psi_L^T G_{ss} = \lambda_c \psi_L^T$). Both left and right eigenvectors as well as λ_c can be efficiently computed by the power iteration method in a similar way as the standard PageRank method. Vectors ψ_R are normalized with $E_s^T \psi_R = 1$ and ψ_L with $\psi_L^T \psi_R = 1$. It is well known (and easy to show) that ψ_L^T is orthogonal to all other right eigenvectors (and ψ_R is orthogonal to all other left eigenvectors) of G_{ss} with eigenvalues different from λ_c. We introduce the operator $\mathcal{P}_c = \psi_R \psi_L^T$ which is the projector onto the eigenspace of λ_c and we denote by $\mathcal{Q}_c = 1 - \mathcal{P}_c$ the complementary projector. One verifies directly that both projectors commute with the matrix G_{ss} and in particular $\mathcal{P}_c G_{ss} = G_{ss}\mathcal{P}_c = \lambda_c \mathcal{P}_c$. Therefore we can derive:

$$(1 - G_{ss})^{-1} = \mathcal{P}_c \frac{1}{1 - \lambda_c} + \mathcal{Q}_c \sum_{l=0}^{\infty} \bar{G}_{ss}^l \qquad (10)$$

with $\bar{G}_{ss} = \mathcal{Q}_c G_{ss} \mathcal{Q}_c$ and using the standard identity $\mathcal{P}_c \mathcal{Q}_c = 0$ for complementary projectors. The expansion in (10) converges rapidly since $\bar{G}_{ss}^l \sim |\lambda_{c,2}|^l$ with $\lambda_{c,2}$ being the second largest eigenvalue which is significantly lower than unity.

The combination of (8) and (10) provides an explicit algorithm feasible for a numerical implementation for modest values of N_r, large values of N_s and of course if sparse matrices G, G_{ss} are considered. We refer the reader to [20] for more advanced implementation considerations.
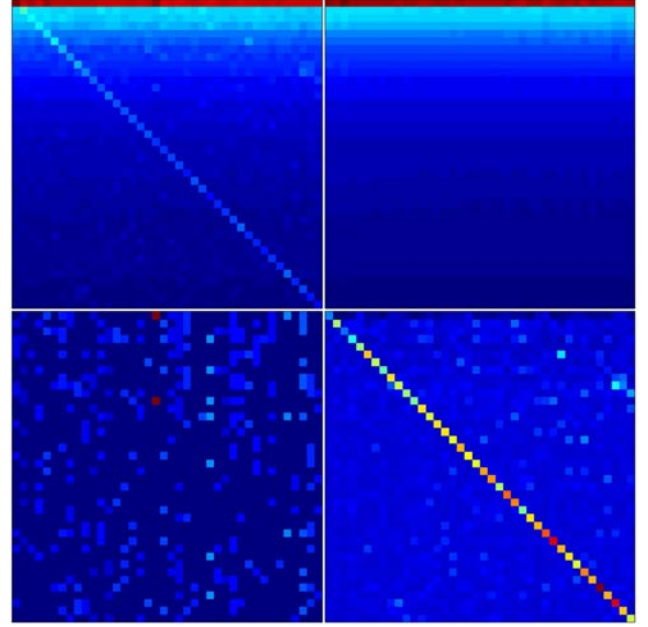

Fig. 3. Density plots of matrices G_R (top left), G_{pr} (top right), G_{rr} (bottom left) and G_{qr} (bottom right) for the reduced network of 40 countries in the EnWiki network. The nodes N_r are ordered in lines by increasing PageRank index (left to right) and in columns by increasing PageRank index from top to bottom. Color scale represents maximum values in red (0.15 in top panels; 0.01 in bottom left panel; 0.03 in bottom right panel), intermediate in green and minimum (approximately zero) in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.4. Decomposition of G_R

On the basis of equations (8)–(10), the reduced Google matrix can be presented as a sum of three components:

$$G_R = G_{rr} + G_{pr} + G_{qr}, \qquad (11)$$

with the first component G_{rr} given by direct matrix elements of G among the selected N_r nodes. The second projector component G_{pr} is given by:

$$G_{pr} = G_{rs} \mathcal{P}_c G_{sr}/(1 - \lambda_c), \quad \mathcal{P}_c = \psi_R \psi_L^T \quad . \qquad (12)$$

The third component G_{qr} is of particular interest in this study as it characterizes the impact of indirect or hidden links. It is given by:

$$G_{qr} = G_{rs}[\mathcal{Q}_c \sum_{l=0}^{\infty} \bar{G}_{ss}^l]G_{sr}, \quad \mathcal{Q}_c = 1 - \mathcal{P}_c,$$
$$\bar{G}_{ss} = \mathcal{Q}_c G_{ss} \mathcal{Q}_c. \qquad (13)$$

We do characterize the strength of these 3 components by their respective weights W_{rr}, W_{pr}, W_{qr} given respectively by the sum of all matrix elements of G_{rr}, G_{pr}, G_{qr} divided by N_r. By definition we have $W_{rr} + W_{pr} + W_{qr} = 1$.

Reduced Google matrix has been computed, together with its components G_{rr}, G_{pr} and G_{qr}, for the English language edition of Wikipedia (EnWiki) and for the $N_R = 40$ countries listed in Table 1. These countries are the ones with top PageRank K in the network of EnWiki. Density plots of G_R, G_{rr}, G_{pr} and G_{qr}, are given in Fig. 3. Countries are ordered by increasing K value. The weight of the three matrix components of G_R are $W_{pr} = 0.96120$, $W_{qr} = 0.029702$ and $W_{rr} = 0.009098$. Predominant component is clearly G_{pr} but as we will explain next, it is not the most meaningful.

The meaning of G_{rr} is clear as it is directly extracted from the global Google matrix G. It gives the direct links between the selected nodes and more specifically the probability $G_{rr}(i, j)$ for the surfer to go directly from column j country to line i country. However, since each column is normalized by the number of outgoing links, absolute probabilities cannot be compared to each other across columns.

The sum of G_{pr} and G_{qr} represents the contribution of all indirect links through the scattering matrix G_{ss}. As seen on Fig. 3, the projector component G_{pr} is composed of nearly identical columns. Moreover, values of each column are proportional to the PageRank of the countries (lines and columns are ordered by increasing K values). As detailed in [20], we observe numerically that $G_{pr} \approx P_r E_r^T /(1 - \lambda_c)$, meaning that each column is close to the normalized vector $P_r/(1 - \lambda_c)$. As such, G_{pr} transposes essentially in G_R the contribution of the first eigenvector of G. We can conclude that even if the overall column sums of G_{pr} account for ∼95–97% of the total column sum of G_R, G_{pr} doesn't offer innovative information compared to PageRank analysis.

A way more interesting contribution is the one of G_{qr}. This matrix captures higher-order indirect links between the N_r nodes due to their interactions with the global network environment. We will refer to these links as *hidden links*. We note that G_{qr} is composed of two parts $G_{qr} = G_{qrd} + G_{qrnd}$ where the first term gives only the diagonal part of the matrix G_{qrd} and thus represents the probabilities to stay on the same node during multiple iterations of $\bar{G}_{ss}$ in (13) while the second matrix captures only non-diagonal terms in G_{qrnd}. As such, G_{qrnd} represents indirect (hidden) links between the N_r nodes appearing via the global network. We note that certain matrix elements of G_{qr} can be negative, which is possible due to the negative terms in $\mathcal{Q}_c = \mathbf{1} - \mathcal{P}_c$ appearing in (13). The total weight of negative elements is however much smaller than W_{qr} (at least 6 times smaller and even non-existing in ArWiki). Of course, the full reduced Google matrix G_R has only positive or zero matrix elements. In the following, our study concentrates mainly on the meaning of G_{qrnd}, emphasizing the meaning of its largest positive values in the study of Section 4.

3. Matrices of world countries

Our study focuses on the networks representing 5 different Wikipedia editions[1] from the set of 24 analyzed in [13]: EnWiki, ArWiki, RuWiki, DeWiki and FrWiki that contain 4.212, 0.203, 0.966, 1.533 and 1.353 million of articles each.

3.1. Selected countries

The 40 countries listed in Table 1 have been selected from the EnWiki network after computing the PageRank for the complete network. The 40 countries with largest PageRank probability have been chosen, and ordered by a local PageRank index K varying between 1 and 40. The most influential countries are found to be at the top values $K = 1, 2, \dots$. In addition we determine the local CheiRank index K^* of the selected countries using the CheiRank vector of the global network [14,21]. At the top of K^* we have the most communicative countries. Table 1 lists K and K^* for EnWiki, ArWiki and RuWiki. Not surprisingly, the order of top countries changes with respect to the edition (for instance, the top country for K is US except for RuWiki whose top country is Russia).

Countries that belong to the same region or having a common piece of history may probably exhibit stronger interactions in Wikipedia. As such, we have created a color code that groups together countries that either belong to the same geographical

region (e.g. Europe, Latin America, Middle East, North-East Asia, South-East Asia) or share a big part of history (former USSR; English speaking countries that are the legacy of the former British Empire). Color code can be seen either in Fig. 1 or in the text color of Table 1.

It is convenient as well to plot all nodes in the (K, K^*) plane to highlight the countries that are the most influential ($K = 1, 2, \dots$) and the most communicative ($K^* = 1, 2, \dots$) at the same time. Fig. 2 plots all 40 countries in the (K, K^*) plane for EnWiki, ArWiki and RuWiki editions. This plot is a bi-objective plot where K and K^* are to be minimized concurrently. It is interesting to look at the set of non-dominated countries which are the ones such that there is no other country beating them for both K and K^*. Cultural bias is obvious here as for EnWiki, this set is composed of {US, JP, IR, AR}, for ArWiki of {US, EG, IT} and for RuWiki of only Russia. We note that according to Fig. 2 some countries with high K value (relative few in-degree and low PageRank probability) act as important diffusers of content (low K^*) even for the language editions being different from the language spoken in those countries (BR, AR, RO). Also countries with low K (having many citations) are relatively poor diffusers (FR, CA). Thus Fig. 2 demonstrates that different cultures attribute different degree of country popularity (PageRank probability and K index) and different communicative degree (CheiRank probability and K^* index). This indicates the nontrivial features of cultures propagation and interactions.

3.2. Density plots of G_R, G_{rr} and G_{qrnd}

For the three EnWiki, ArWiki and RuWiki editions, Fig. 4 plots the density of matrices G_R, G_{qrnd} and G_{rr}. We keep for all plots the same order of countries extracted from the EnWiki network. This is meant to highlight cultural differences among Wikipedia editions. In the G_R plots, it is clear that this matrix is dominated by the projector G_{pr} contribution, which is proportional to the global PageRank probabilities. In G_R, not surprisingly, the cultural bias is pronounced due to its strong tie to PageRank. For instance, Egypt is much more important in ArWiki than in other editions, and Russia is the top country in RuWiki.

The information from direct links between countries is provided by G_{rr}. As expected, the per-column normalization prevents a meaningful per-line analysis. For EnWiki and ArWiki, the respective columns of Mexico and Hungary are predominant due to their little number of outgoing links. On the contrary, G_{qrnd} offers a much more unified view of countries interactions as it seems to highlight more general interaction that are less biased by cultural views. For instance, for these three Wikipedia editions, the hidden links connecting Taiwan to China and Pakistan to India are really strong in G_{qrnd}. The link connecting Ukraine to Russia is very strong in EnWiki and ArWiki. It is surprisingly absent from RuWiki (or maybe this is exactly a cultural bias we are observing since during a long time both countries were part of USSR, and there was thus no specific difference between them). Other interesting hidden links are highlighted in EnWiki as New-Zealand is directed to Australia or in RuWiki linking Canada to the USA.

3.3. Friends and followers

In order to better capture the interactions provided by G_{qrnd} and G_R, we have listed for all 5 Wikipedia editions the top 4 friends and top 4 followers of a set of 7 leading countries. One leading country per group has been selected: US for English speaking countries, Brazil for Latin America, France for Europe, Japan for North-East Asia, India for South-East Asia, Russia for the Soviet block and Turkey for the Middle-East. To pick them inside each group, we have chosen the country whose worst PageRank order over all 5 Wikipedia editions is the highest.
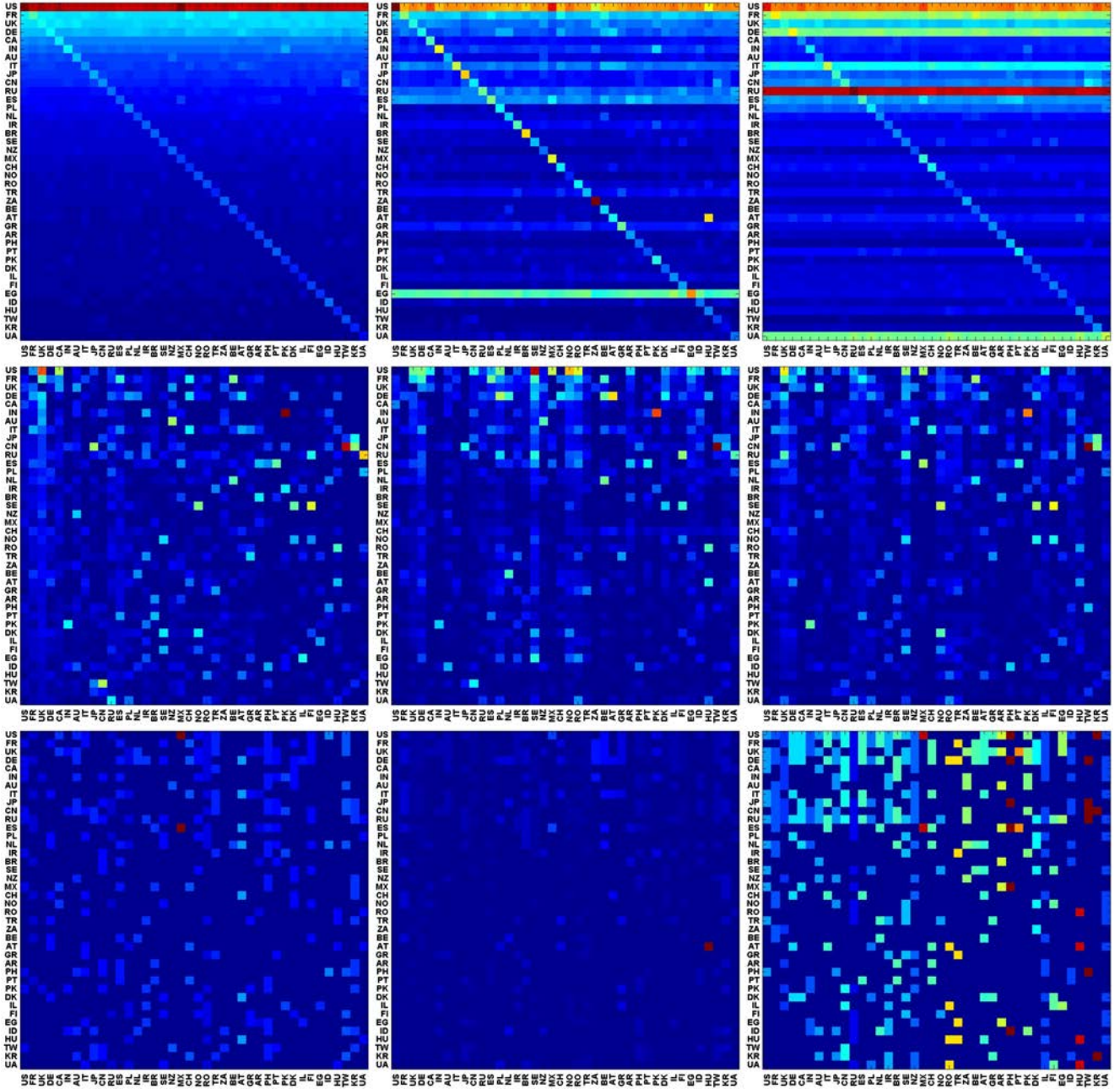
---

[1] Data collected mid February 2013.

Fig. 4. Density plots of the matrix elements of G_R (first line), G_{qrnd} (second line) and G_{rr} (third line) for the reduced network of 40 countries of EnWiki (left column), ArWiki (middle column) and RuWiki (right column). The country names are given on the axes in the order names in Table 1, thus the nodes N_r are ordered in lines and columns by the reference PageRank of EnWiki. The colors represent maximum (red corresponds to: 0.15, 0.19, 0.13 in top panels from left to right; 0.01, 0.03, 0.012 in middle panels and 0.01, 0.011, 0.006 in bottom panels respectively), intermediate (green) and minimum (blue for zero) values for a given matrix. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For each leading country, we extract from both matrices G_{qrnd} and G_R the top 4 *Friends* (resp. *Followers*) of country j given by the 4 best values of the elements of column j (resp. of line j). In other words, it corresponds to destinations of the 4 strongest outgoing links of j and the countries at the origin of strongest 4 ingoing links. A summary of relevant results is given in Table 2 to show cross-culture interactions.

Looking at G_R friends, top friends of leading countries are strongly related to the top PageRank countries (we have a predominance of US, France and Germany). Similarly, cross-edition followers of US are Mexico and Canada, and followers of Japan are China, Korea and Taiwan. On the opposite, higher-order interactions of G_{qrnd} are not as much influenced by PageRank. More

subtle but realistic interactions appear: Canada is always identified as a hidden friend of USA while it was never the case in G_R. Similarly, Ukraine is always tagged as a hidden friend of Russia; Italy and Spain as friends of France. Thus G_{qrnd} seems to emphasize more fine-grained regional interactions. Next section exploits the concept of friends and followers to create new network representations derived from G_R and G_{qrnd}.

4. Networks of 40 countries

This study concentrates again on the same 7 leading countries as before. Top 4 friends and top 4 followers of these leading countries are extracted from G_R and G_{qrnd} to plot the graphs of Figs. 5

Table 2

Cross-edition direct friends and followers extracted from G_R matrix (top table) and from G_{qrnd} (bottom table) for the top countries of each area. For each top country, we list the direct friends (followers) present in the direct friends list given by all five Wikipedia editions, the ones present in 4 editions out of 5 and the ones present in 3 editions out of five.

Top country	G_R Wiki friends present in			G_R Wiki followers present in		
	all 5 editions	4 out of 5 editions	3 out of 5 editions	all 5 editions	4 out of 5 editions	3 out of 5 editions
US	FR	DE	UK	MX–CA	IL	UK
BR	US–FR	DE		AR–PT		MX
FR	US	DE		BE	ES	IT
RU	US–FR	DE		UA–FI		PL
TR	US–FR	DE		GR–IR		
JP	US–FR	DE		KR–TW–CN	PH	
IN	US–FR		DE	PK–ID		ZA

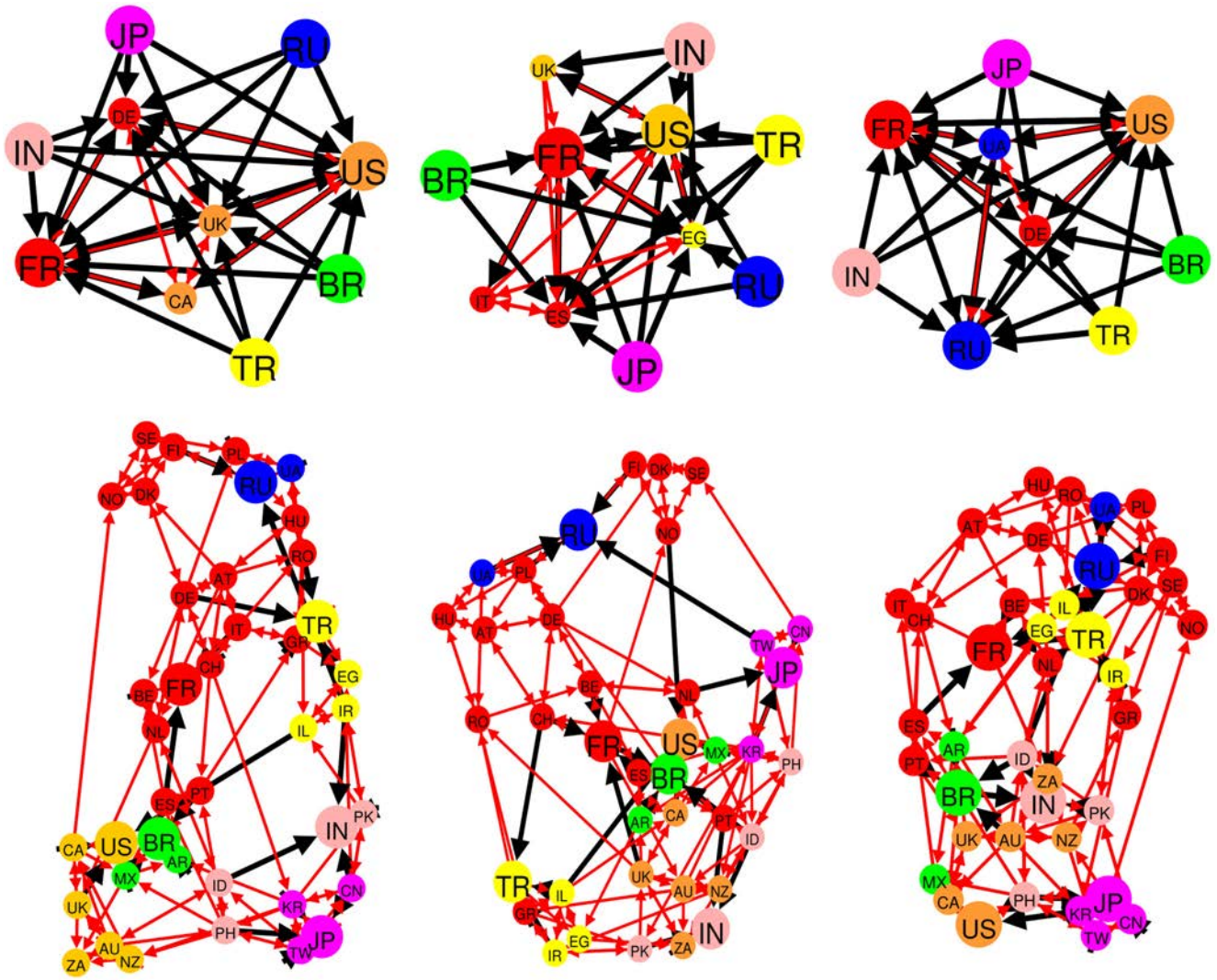| Top country	G_{qr} Wiki friends present in			G_{qr} Wiki followers present in		
	all 5 editions	4 out of 5 editions	3 out of 5 editions	all 5 editions	4 out of 5 editions	3 out of 5 editions
US	CA	UK–MX	FR		UK	CA–MX
BR	PT–AR	ES–MX		AR	PT	UK
FR	IT–DE		US–BE		UK–BE–ES	CH
RU	UA	DE	CN–PL–US	FI	UA–PL	IR
TR	GR	IR–RU		GR–IR	EG–RO	
JP		CN–TW–US–KR		KR–CN–TW		
IN	PK–CN	ID	IR	PK–ID		CN

Fig. 5. Network structure of friends (top line) and followers (bottom line) induced by the 7 top countries of each geographical area (US, FR, IN, JP, BR, TR, RU) in G_R. Results are plotted for EnWiki (left column), ArWiki (middle column) and RuWiki (right column). Node colors represent geographic appartenance to a group of countries (cf. Table 1 for details). Top (bottom) graphs: the top country node points (is pointed by) with a bold black arrow to its top 4 friends (followers). Red arrows show friends of friends (resp. followers of followers) interactions computed until no new edges are added to the graph. Drawn with [22]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
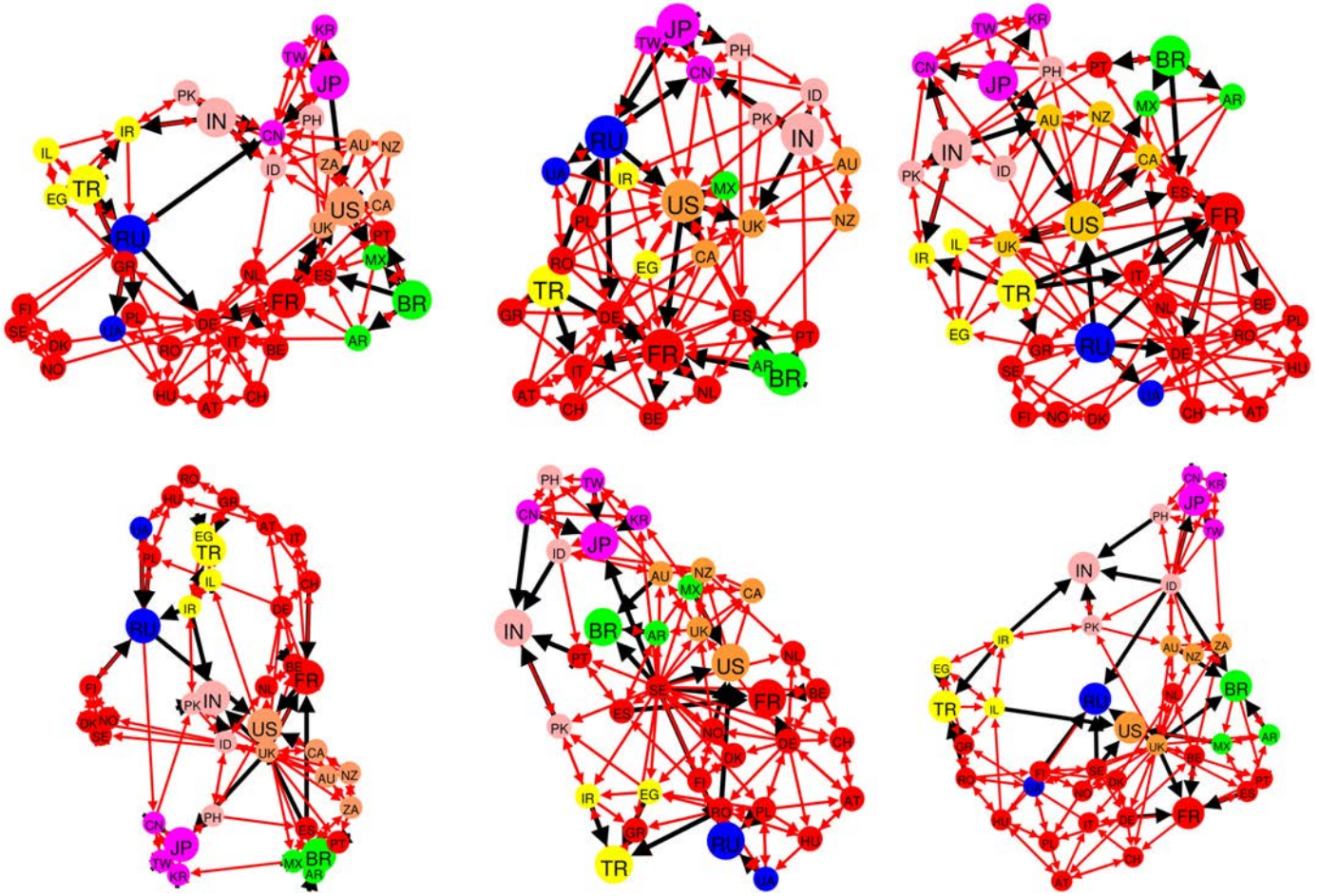
Fig. 6. Same legend as Fig. 5 except friends and followers are computed from G_{qrnd} (hidden relationships). (For interpretation of the colors in this figure, the reader is referred to the web version of this article.)

and 6, respectively. Note that Fig. 6 essentially highlights hidden links. The black thick arrows identify the top 4 friends and top 4 followers interactions. Red arrows represent the friends of friends (respectively the followers of followers) interactions that are computed recursively until no new edge is added to the graph. All graphs are plotted using a force direct layout.

The networks of friends obtained from G_R never expand to the full set of 40 nodes. They only concentrate on about 10 countries (including the 7 leading ones). This happens since G_R is dominated by the projector component. Looking at the follower graphs, more information can be observed. North-East Asian, Middle-Eastern and Latin American create, in all editions, a cluster of nodes densely interconnected. European countries enclose Russia and Ukraine as these countries are linked to EU countries that were part of the former Soviet Union zone of influence (e.g. Romania, Hungary, Finland, etc.). The networks of followers end up almost spanning the full set of 40 countries.

The networks of friends obtained from G_{qrnd} don't concentrate to a limited set of countries as it is the case for G_R. They end up spanning the full set of countries. The hidden friend links show that the interactions between the geographical groups are coherent. North-East Asian countries are linked to South-East Asian countries and to English speaking countries in EnWiki and RuWiki. Interestingly, the set of Baltic countries (SE, NO, DK, FI) create most of the time full meshes, and interconnect Europe and Russia. Cultural bias can be observed as well in these plots. For non-Arabic editions, Middle-Eastern countries create a well-connected cluster

of nodes. But for ArWiki, Turkey exhibits a stronger connection with Europe than with the other Middle-Eastern countries. In the view of Arabic countries, Turkey is seen closer to Europe than others for sure.

5. Conclusion

This work offers a new perspective for future geopolitics studies. It is possible to extract from multi-cultural Wikipedia networks a global understanding of the interactions between countries at a global, continental or regional scale. On the basis of the reduced Google matrix analysis of the Wikipedia network we determined the relations of friends and followers of countries on purely statistical mathematical grounds obtained from the huge knowledge accumulated by Wikipedia editions in different languages. This approach provides results independent of cultural bias. The reduced Google matrix theory has been shown to capture hidden and indirect interactions among countries, resulting in new knowledge on geopolitics.

References

[1] E. Jones, The European Miracle: Environments, Economies and Geopolitics in the History of Europe and Asia, Cambridge University Press, Cambridge, MA, 2003.

[2] J. Dittmer, K. Dodds, Popular geopolitics past and future: fandom, identities and audiences, Geopolitics 13 (2008) 3.

[3] http://www.wikipedia.org (accessed August 2016).

[4] Encyclopaedia Britannica, http://www.britannica.com/ (accessed August 2016).

[5] J. Giles, Nature 438 (2005) 900.

[6] J.M. Reagle Jr., Good Faith Collaboration: The Culture of Wikipedia, MIT Press, Cambridge, MA, 2010.

[7] F.A. Nielsen, Wikipedia research and tools: review and comments, available at SSRN: http://dx.doi.org/10.2139/ssrn.2129874, 2012.

[8] S. Brin, L. Page, Comput. Netw. ISDN Syst. 30 (1998) 107.

[9] A.M. Langville, C.D. Meyer, Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press, Princeton, 2006.

[10] A.O. Zhirov, O.V. Zhirov, D.L. Shepelyansky, Eur. Phys. J. B 77 (2010) 523.

[11] Y.-H. Eom, K.M. Frahm, A. Benczur, D.L. Shepelyansky, Eur. Phys. J. B 86 (2013) 492.

[12] Y.-H. Eom, D.L. Shepelyansky, PLoS ONE 8 (10) (2013) e74554.

[13] Y.-H. Eom, P. Aragon, D. Laniado, A. Kaltenbrunner, S. Vigna, D.L. Shepelyansky, PLoS ONE 10 (3) (2015) e0114825.

[14] L. Ermann, K.M. Frahm, D.L. Shepelyansky, Rev. Mod. Phys. 87 (2015) 1261.

[15] J. Lages, A. Patt, D.L. Shepelyansky, Eur. Phys. J. B 89 (2016) 69.

[16] M.H. Hart, The 100: Ranking of the Most Influential Persons in History, Citadel Press, N.Y., 1992.

[17] Academic ranking of world universities, http://www.shanghairanking.com/ (accessed August 2016).

[18] R. Meusel, S. Vigna, O. Lehmberg, C. Bizer, J. Web Sci. 1 (2015) 33.

[19] K.M. Frahm, D.L. Shepelyansky, arXiv:1602.02394 [physics.soc], 2016.

[20] K.M. Frahm, K. Jaffrès-Runser, D.L. Shepelyansky, Eur. Phys. J. B 89 (2016) 269.

[21] A.D. Chepelianskii, arXiv:1003.5455 [cs.SE], 2010.

[22] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: Proc. of International AAAI Conference on Weblogs and Social Media, 2009.