Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

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*

ARTICLE INFO

Article history:
Received 23 December 2016
Received in revised form 9 June 2017
Accepted 12 June 2017
Available online xxxx
Communicated by P.R. Holland

Keywords:
Markov chains
Google matrix
Wikipedia network
Geopolitical interactions of countries

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).

http://dx.doi.org/10.1016/j.physleta.2017.06.021
0375-9601/© 2017 Elsevier B.V. All rights reserved.

Please cite this article in press as: K.M. Frahm et al., Multi-cultural Wikipedia mining of geopolitics interactions leveraging reduced Google matrix analysis, Phys. Lett. A (2017), http://dx.doi.org/10.1016/j.physleta.2017.06.021

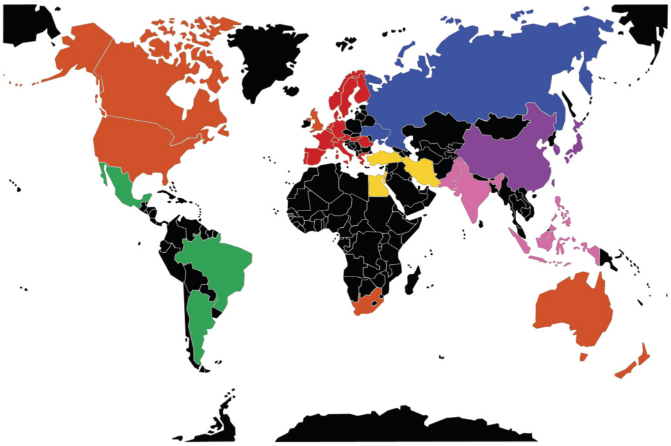

Fig. 1. Geographical distribution of the 40 selected countries. Color code groups countries into 7 sets: orange (OC) for English speaking countries, blue (BC) for former Soviet union ones, red (RC) for European ones, green (GC) for Latin American ones, yellow (YC) for Middle Eastern ones, purple (PUC) 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.)

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.

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

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.