EPJ manuscript No.
(will be inserted by the editor)

Wikipedia Ranking of World Universities

José Lages[1a], Antoine Patt[1], and Dima L. Shepelyansky[2b]

[1] Institut UTINAM, Observatoire des Sciences de l'Univers THETA, CNRS, Université de Franche-Comté, 25030 Besançon, France

[2] Laboratoire de Physique Théorique du CNRS, IRSAMC, Université de Toulouse, UPS, 31062 Toulouse, France

Received: 29 November 2015 / Revised version: date

Abstract. We use the directed networks between articles of 24 Wikipedia language editions for producing the Wikipedia Ranking of World Universities (WRWU) using PageRank, 2DRank and CheiRank algorithms. This approach allows to incorporate various cultural views on world universities using the mathematical statistical analysis independent of cultural preferences. The Wikipedia ranking of top 100 universities provides about 60 percent overlap with the Shanghai university ranking demonstrating the reliable features of this approach. At the same time WRWU incorporates all knowledge accumulated at 24 Wikipedia editions giving stronger highlights for historically important universities leading to a different estimation of efficiency of world countries in university education. The historical development of university ranking is analyzed during ten centuries of their history.

PACS. 89.75.Fb Structures and organization in complex systems – 89.75.Hc Networks and genealogical trees – 89.20.Hh World Wide Web, Internet

1 Introduction

According to the UNESCO reports the higher education is definitely at the heart of modern society development and related academic revolution (see e.g. [1]). Thus the analysis of the efficiency of university education in different countries becomes of political importance for the country future development. One of the important tools of this analysis is the university ranking reviewed in high details at [2]. Indeed, it is now well established that the Academic Ranking of World Universities (ARWU), compiled by Shanghai Jiao Tong University since 2003 (Shanghai ranking) [3], produced a significant impact on evaluation of national universities both on educational and political levels [1,2]. Thus, for example, ARWU affected the French strategies LABEX, IDEX in high education [4]. Also the Russian Academic Excellence Project with significant financial investments [5] in many respects has been initiated by ARWU. Other examples are reviewed in [2]. At present there are several additional university rankings which are based on various evaluation methods of university efficiency in research and education (see e.g. [6,7,8]).

The scientific analysis of strong and weak features of various university ranking methods is performed by various research groups as reported for example in [9,10,11, 12]. A comparative analysis of various approaches is given in [2,13]. It is in general accepted that the world university rankings play an important role for development of higher education in the world countries, even if there are various opinions about each approach.

The above scientific studies definitely show the importance of university ranking. These ranking approaches are based on human selection rules which can not be complete or can favor certain cultural choices and preferences. Thus it is useful to have an independent mathematical statistical method which would rank universities independently of any human rules. Such a method has been proposed in [14] being based on the mathematical analysis of the human knowledge accumulated at English Wikipedia by year 2009 (www.wikipedia.org). This approach is based on a directed network of citations between all available articles of Wikipedia, construction of the corresponding Markov chain transitions [15] and the Google matrix G, introduced by Brin and Page in 1998 [16] for hypertext analysis of the World Wide Web (WWW). The construction rules of G matrix and description of its spectral properties for various directed networks are given in [17,18]. The general scale-free properties of complex networks are described in [19]. The studies performed in [14,20] demonstrated that this approach recovers about 70% and 80% of top 100 and top 10 universities of ARWU and that this overlap remains stable during the time evolution of English Wikipedia during the years $2004-2011$.

A similar approach based on the Wikipedia network was used for ranking of historical figures of English Wikipedia [14,20]. The extension of this approach to 9 [21] and 24 language editions of Wikipedia [22] allowed to take into account various cultural view points and improve the

[a] *email address:* jose.lages@utinam.cnrs.fr
[b] *email address:* dima@irsamc.ups-tlse.fr

overlap of top 100 historical figures from Wikipedia with the Hart top 100 people, who according to him, most influenced human history [23]. The approaches of different groups to the Wikipedia ranking of historical figures are discussed in [14,22,24]. The results for the top 100 historical figures of Wikipedia approve the validity of this mathematical ranking approach based on human knowledge accumulated in various language editions of Wikipedia.

In this work we extend this approach creating the Wikipedia Ranking of World Universities (WRWU). We use the network data set of 24 Wikipedia language editions collected at [22]. These 24 languages cover 59% of world population and 68% of the total number of Wikipedia articles in all 287 languages. On the basis of the developed analysis we determine the most influential universities in the world and consider their time and geographical evolution on a scale of 10 centuries of human history. This study also allows to consider the various cultural preferences in the importance of concrete universities by different countries. Our WRWU results have about 60% and 90% overlap with the top 100 and top 10 list of ARWU.

The paper is constructed as follows: In Section 2, we describe Wikipedia data sets used in this work and we introduce the WRWU approach which is based on the Google matrix and PageRank, CheiRank, 2DRank algorithms. In Section 3, results of WRWU are compared to ARWU with a geographical and temporal analysis. In Section 4, we study entanglement of cultures and their interactions through WRWU results. Finally, the discussion of the results is given in Section 5.

2 Description of data sets and methods

We consider 24 Wikipedia language editions already used to rank historical figures of Wikipedia [22]: Arabic (AR), Danish (DA), German (DE), Greek (EL), English (EN), Spanish (ES), Persian (FA), French (FR), Hebrew (HE), Hindi (HI), Hungarian (HU), Italian (IT), Japanese (JA), Korean (KO), Malaysian (MS), Dutch (NL), Polish (PL), Portuguese (PT), Russian (RU), Swedish (SV), Thai (TH), Turkish (TR), Vietnamese (VI), Chinese (ZH). Titles of Wikipedia articles and hyperlinks between articles were collected in middle February 2013 (see [22] for data preparation details).

2.1 Network definition

Following [22], we consider each of the Wikipedia language editions as an isolated directed network whose nodes are articles and the directed links are formed by citations from one article to another article. In this study we do not consider hyperlinks between different language editions. We associate to a given network an adjacency matrix A with elements A_{ij} being 1 if node (article) j points towards node (article) i and 0 otherwise. A network associated to a given Wikipedia language edition containing N articles connected with N_ℓ hyperlinks is then characterized by its

Table 1. Wikipedia directed networks from 24 considered language editions; here N is the number of articles. Wikipedia data were collected in middle February 2013 [22].

Edition	Language	N	Edition	Language	N
EN	English	4212493	VI	Vietnamese	594089
DE	German	1532978	FA	Persian	295696
FR	French	1352825	HU	Hungarian	235212
NL	Dutch	1144615	KO	Korean	231959
IT	Italian	1017953	TR	Turkish	206311
ES	Spanish	974025	AR	Arabic	203328
RU	Russian	966284	MS	Malaysian	180886
PL	Polish	949153	DA	Danish	175228
JA	Japanese	852087	HE	Hebrew	144959
SV	Swedish	780872	HI	Hindi	96869
PT	Portuguese	758227	EL	Greek	82563
ZH	Chinese	663485	TH	Thai	78953

$N \times N$ adjacency matrix A containing N_ℓ non zero A_{ij} elements. The parameters of the networks constructed from 24 Wikipedia language editions are given in Table 1 (see also [22]). The country codes (CC) and language codes (LC) are given in Table 2. The CC codes follow ISO 3166-1 alpha-2 standard [25] and the LC codes are language edition codes of Wikipedia, the code WR represents all languages other than the considered 24 languages.

2.2 Google matrix

We suppose that a random surfer hops from a node j to any connected node i ($A_{ij} = 1$) with probability $1/k_{out}(j)$ where $k_{out}(j) = \sum_{i=1}^{N} A_{ij} \neq 0$ is the node j out-degree, *i.e.* the number of links from node j to other nodes. If node j is a dangling node without outgoing links ($k_{out}(j) = 0$), then we assume that a random surfer hops to any of the network nodes N with the probability $1/N$. Then the matrix of Markov transitions S is defined by its elements $S_{ij} = A_{ij}/k_{out}(j)$ if $k_{out}(j) \neq 0$ and $S_{ij} = 1/N$ otherwise. The Google matrix G is defined by the standard relation [16,17]:

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

where α is the damping factor. We use throughout the paper the conventional value $\alpha = 0.85$. The values α in the range $0.5 \leq \alpha < 0.95$ do not affect the ranking [17,18,22].

2.3 PageRank, CheiRank and 2DRank algorithms

The PageRank algorithm [16,17] allows to rank all nodes of the network. Let us assume that a random surfer journey starts from node k. We define $P_i(t)$ the probability that a random surfer reaches node i after t iterations with $P_k(0) = 1$ and $P_i(0) = 0$ for $i \neq k$. The probability vector $\mathbf{P}(t)$ whose the components are the probabilities $P_i(t)$ is given by

$$\mathbf{P}(t) = \underbrace{GG\ldots G}_{t \text{ times}} \mathbf{P}(0) = G^t \mathbf{P}(0). \qquad (2)$$

Table 2. List of countries with corresponding country codes (CC) and language codes (LC). Only countries appearing in the top 100 universities of 24 Wikipedia editions using PageRank, CheiRank, and 2DRank algorithms are listed here. LC is determined by the most spoken language in the given country. Country codes (CC) follow ISO 3166-1 alpha-2 standard [25]. Language codes are based on language edition codes of Wikipedia; WR represents all languages other than the considered 24 languages. The data are represented by three columns with CC Country LC.

CC	Country	LC	CC	Country	LC	CC	Country	LC
AE	United Arab Emirates	AR	GU	Guam	EN	OM	Oman	AR
AF	Afghanistan	FA	GY	Guyana	EN	PA	Panama	ES
AL	Albania	WR	HK	Hong Kong	ZH	PE	Peru	ES
AM	Armenia	WR	HN	Honduras	ES	PG	Papua New Guinea	EN
AO	Angola	PT	HR	Croatia	WR	PH	Philippines	EN
AR	Argentina	ES	HT	Haiti	FR	PK	Pakistan	HI
AT	Austria	DE	HU	Hungary	HU	PL	Poland	PL
AU	Australia	EN	ID	Indonesia	WR	PR	Puerto Rico	ES
AZ	Azerbaijan	TR	IE	Ireland	EN	PS	State of Palestine	AR
BD	Bangladesh	WR	IL	Israel	HE	PT	Portugal	PT
BE	Belgium	NL	IN	India	HI	PY	Paraguay	ES
BF	Burkina Faso	FR	IQ	Iraq	AR	QA	Qatar	AR
BG	Bulgaria	WR	IR	Iran	FA	RO	Romania	WR
BH	Bahrain	AR	IS	Iceland	WR	RS	Serbia	WR
BJ	Benin	FR	IT	Italy	IT	RU	Russia	RU
BN	Brunei	MS	JM	Jamaica	EN	RW	Rwanda	EN
BR	Brazil	PT	JO	Jordan	AR	SA	Saudi Arabia	AR
BS	Bahamas	EN	JP	Japan	JA	SD	Sudan	AR
BT	Bhutan	WR	KE	Kenya	EN	SE	Sweden	SV
BY	Belarus	RU	KG	Kyrgyzstan	WR	SG	Singapore	ZH
CA	Canada	EN	KH	Cambodia	WR	SI	Slovenia	WR
CF	Central African Republic	FR	KM	Comoros	FR	SK	Slovakia	WR
CH	Switzerland	DE	KP	North Korea	KO	SO	Somalia	WR
CI	Ivory Coast	FR	KR	South Korea	KO	SR	Suriname	NL
CL	Chile	ES	KW	Kuwait	AR	SV	El Salvador	ES
CN	China	ZH	KZ	Kazakhstan	WR	SY	Syria	AR
CO	Colombia	ES	LA	Laos	WR	SZ	Swaziland	EN
CR	Costa Rica	ES	LB	Lebanon	AR	TH	Thailand	TH
CU	Cuba	ES	LK	Sri Lanka	WR	TJ	Tajikistan	WR
CY	Cyprus	EL	LR	Liberia	EN	TL	Timor-Leste	PT
CZ	Czech Republic	WR	LT	Lithuania	WR	TN	Tunisia	AR
DE	Germany	DE	LV	Latvia	WR	TR	Turkey	TR
DK	Denmark	DA	LY	Libya	AR	TW	Taiwan	ZH
DO	Dominican Republic	ES	MA	Morocco	AR	TZ	Tanzania	WR
DZ	Algeria	AR	MC	Monaco	FR	UA	Ukraine	WR
EC	Ecuador	ES	MD	Moldova	WR	UG	Uganda	EN
EE	Estonia	WR	MK	Macedonia	WR	UK	United Kingdom	EN
EG	Egypt	AR	MM	Myanmar	WR	US	United States	EN
ES	Spain	ES	MN	Mongolia	WR	UY	Uruguay	ES
ET	Ethiopia	EN	MT	Malta	EN	UZ	Uzbekistan	WR
FI	Finland	WR	MW	Malawi	EN	VA	Holy See	IT
FJ	Fiji	EN	MX	Mexico	ES	VE	Venezuela	ES
FO	Faroe Islands	DA	MY	Malaysia	MS	VN	Vietnam	VI
FR	France	FR	NG	Nigeria	EN	YE	Yemen	AR
GE	Georgia	WR	NL	Netherlands	NL	ZA	South Africa	WR
GH	Ghana	EN	NO	Norway	WR	ZW	Zimbabwe	EN
GL	Greenland	DA	NP	Nepal	WR			
GR	Greece	EL	NZ	New Zealand	EN			

Providing $\alpha < 1$, for any given $\mathbf{P}(0)$, the probability vector $\mathbf{P}(t)$ converges towards an unique stationary vector $\mathbf{P}$ as the number of iterations increases. This is the right eigenvector of G matrix with the eigenvalue $\lambda = 1$ ($G\mathbf{P} = \lambda\mathbf{P}$). In our numerical simulations we compute iteratively $\mathbf{P}$ up to a precision of 10^{-17}, *i.e.* we compute $P(t)$ up to iteration t' such as $\sum_{i=1}^{N} |P_i(t') - P_i(t'-1)| \leq 10^{-17}$. The ith component of $\mathbf{P}$, P_i, gives the average proportion of time spent by a random surfer on node i. Ordering the probabilities P_i from biggest to smallest gives the PageRank [16] index K with $K = 1$ ($K = N$) associated to node with maximum (minimum) probability.

It is also useful to consider the network with inverted direction of links. Then the matrix of Markov transitions is defined as $S^*_{ij} = A_{ji}/k_{in}(j)$ if $k_{in}(j) \neq 0$ and $S_{ij} = 1/N$ otherwise. Here $k_{in}(j) = \sum_{j=1}^{N} A_{ij}$ is the node j in-degree *i.e.* the number of links to node j from other nodes. The associated dual Google matrix G^* is consequently $G^*_{ij} = \alpha S^*_{ij} + (1-\alpha)/N$. Similarly to PageRank, it is possible to define a probability vector

$$\mathbf{P}^*(t) = \underbrace{G^* G^* \ldots G^*}_{t \text{ times}} \mathbf{P}^*(0) = {G^*}^t \mathbf{P}^*(0). \qquad (3)$$

which, providing $\alpha < 1$, for any given $\mathbf{P}^*(0)$, converges towards an unique stationary probability vector $\mathbf{P}^*$. Probability P^*_i gives the average time spent on node i by a random surfer evolving on the inverted directed network. The probability vector $\mathbf{P}^*$ is the right eigenvector of the matrix G^* with the eigenvalue $\lambda = 1$ ($G^*\mathbf{P}^* = \lambda\mathbf{P}^*$). The statistical properties of this CheiRank vector have been analyzed in [26] (see also [14,18]). By ordering the probabilities P^*_i from largest to smallest values gives the CheiRank index with $K^* = 1$ ($K^* = N$) associated to node with maximum (minimum) probability.

The probability of PageRank vector is proportional to the number of ingoing links while the probability of the CheiRank vector is proportional to the number of outgoing links (see e.g. [17,18]). It is also possible to define a third ranking, 2DRank, which combines PageRank and CheiRank [14]. Assuming a node with PageRank K and CheiRank K^*, the 2DRank index for this node is $K_2 = \max\{K, K^*\}$. The 2DRank algorithm and 2DRank index K_2 are described in detail at [14]. Thus the PageRank index K have at the top well known articles of Wikipedia (e.g. world countries) while the CheiRank index K^* has at the top very communicative article (e.g. listings of geographical names, prime ministers etc.). The top articles of 2DRank index K_2 are those which are both well known and communicative (see [14,18,26]). We note that PageRank and CheiRank appear very naturally in the trade networks corresponding to import and export flows [18]. For the Wikipedia networks the global properties of PageRank, CheiRank and 2DRank have been discussed in detail in [14,20,21,22].

2.4 Rankings of world universities

For each individual Wikipedia language edition we rank all N articles using PageRank, CheiRank, and 2DRank algorithms. We consequently obtain three different global rank indexes K, K^*, K_2 from which we extract articles devoted to an university or an institution of higher education and research. We extract articles with a title containing the keyword “university” in the corresponding language. Additional extractions with keywords such as *e.g.* ”institute”, ”school”, ”college” have also been performed. A manual *a posteriori* check of the automatic extraction have been done to remove *e.g* fictional universities, colleges and schools of lower education from the list of top 100 universities of each edition. We extract also institutions of higher education and research which are designated by acronyms such as *e.g* ”ETH Zurich”. The organizations of pure research (*e.g.* CNRS, NASA) are not taken into account.

For example in the articles of the French Wikipedia edition, ranked by the PageRank algorithm, the first article of university is entitled “Université Harvard” with

PageRank index $K = 904$, then in the second position comes "École polytechnique (France)" with $K = 1549$, and in the third position comes "Université d'Oxford" with $K = 1558$. Thus the top 3 PageRank universities in French Wikipedia edition are: 1. Harvard University, 2. École polytechnique, and 3. Oxford University. The same procedure is used to rank universities with CheiRank and 2DRank algorithm. In this way we determine the top 100 universities for each of 24 Wikipedia editions. Then each university U obtains associated rank index $1 \leq R_{U,E,A} \leq 100$ corresponding to its position in the top 100 list obtained for edition E by algorithm A.

Following [22], for each type of algorithm, we define a global rank from the rank $R_{U,E,A}$ of each of 24 Wikipedia editions. Let us define the ranking score [22]

$$\Theta_{U,A} = \sum_{E} (101 - R_{U,E,A}) \quad (4)$$

where the summation is done over 24 Wikipedia editions. For a given ranking algorithm, the ranking score Θ of an university will be high if it appears well ranked in various Wikipedia language editions. We use the Θ-score (4) to merge the 24 world universities rankings obtaining the global ranking for all 24 Wikipedia language editions. The largest value of $\Theta_{U,A}$ determines the first top world university, the next gives the second world university etc. Top ranked universities obviously appear in most of the Wikipedia language editions. For each university we also determine the number of appearances $1 \leq N_a \leq 24$ in the top 100 list of universities of each edition. The global WRWU lists for each algorithm are given at the web page [27]. In total there are $N_u = 1025, 1379, 1560$ different universities for PageRank, CheiRank, 2DRank algorithms respectively. We notate these global lists as WPRWU, WCRWU, W2RWU respectively.

The top 100 universities for each edition and each algorithm are given at [27]. To each of these universities we attribute the year of its foundation (century), country of its foundation (corresponding to actual country borders given at [28]) and language corresponding to this country defined in Table 2. From the global ranks WPRWU, WCRWU, W2RWU we obtain local university ranking corresponding to each language (selection of universities of the same language). The top 10 universities from these ranks are given at the web page [27]. We use the indexes K_U and $K^*{}_U$ for ranks of global top 100 universities of WPRWU and WCRWU respectively.

Table 3. List of the first 10 universities of the Wikipedia PageRank of World Universities. The score Θ_{PR} is defined by (4); N_a is the number of appearances of a given university in the top 100 lists Wikipedia editions.

Rank	WPRWU	Θ_{PR}	N_a
1st	University of Cambridge	2272	24
2nd	University of Oxford	2247	24
3rd	Harvard University	2112	22
4th	Columbia University	2025	23
5th	Princeton University	1887	23
6th	Massachusetts Institute of Technology	1869	21
7th	University of Chicago	1783	22
8th	Stanford University	1765	21
9th	Yale University	1716	20
10th	University of California, Berkeley	1557	19

Table 4. List of the first 10 universities of the Wikipedia CheiRank of World Universities, other parameters are as in Table 3.

Rank	WCRWU	Θ_{CR}	N_a
1st	University of Oxford	1191	18
2nd	Harvard University	1025	17
3rd	Yale University	1021	16
4th	Massachusetts Institute of Technology	816	16
5th	University of Cambridge	803	11
6th	Columbia University	779	14
7th	Uppsala University	751	11
8th	University of Göttingen	735	13
9th	Humboldt University of Berlin	703	12
10th	Moscow State University	699	14

Table 5. List of the first 10 universities of the Wikipedia 2DRank of World Universities, other parameters are as in Table 3.

Rank	W2RWU	Θ_{2R}	N_a
1st	University of Cambridge	942	16
2nd	Columbia University	786	11
3rd	Stanford University	712	11
4th	Harvard University	683	11
5th	Yale University	609	11
6th	Princeton University	596	11
7th	Massachusetts Institute of Technology	581	10
8th	Humboldt University of Berlin	578	10
9th	Nanjing University	516	8
10th	Johns Hopkins University	511	9

3 WRWU results

We discuss here the results of the WRWU obtained by the methods described in the previous Section. The WRWU results are compared with the ARWU top 100 list taken for year 2013 thus corresponding to the dating of considered Wikipedia networks. We also analyze the WRWU in dependence on the university foundation century and consider the geographical distribution of top universities of WRWU. The tables of the top 10 universities of WPRWU, WCRWU, W2RWU, ARWU are presented in Tables 3, 4, 5, 6. These tables have the well known world university. We discuss the properties of WRWU in more detail in next subsections.

3.1 Comparison of WRWU and ARWU

According to the Tables 3,4,5,6 the overlaps η_{10} (fraction of same names among top 10) with the top 10 list of ARWU are $\eta_{10} = 0.9, 0.5, 0.6$ respectively for WPRWU, WCRWU, W2RWU. Thus WPRWU gives a reliable ranking of world universities being close to the choice of ARWU. However, at the top positions WPRWU places Cambridge,

Table 6. List of the first 10 universities of ARWU 2013 [3]. The 3 last columns show the difference between ARWU rank and WPRWU, WCRWU, W2RWU ranks.

Rank	ARWU	WPRWU	WCRWU	W2RWU
1st	Harvard University	-2	-1	-3
2nd	Stanford University	-6	-9	-1
3rd	University of California, Berkeley	-7	-17	-13
4th	Massachusetts Institute of Technology	-2	0	-3
5th	University of Cambridge	+4	0	+4
6th	California Institute of Technology	-22	-71	-124
7th	Princeton University	+2	-15	+1
8th	Columbia University	+4	+2	+6
9th	University of Chicago	+2	-45	-70
10th	University of Oxford	+8	+9	-2

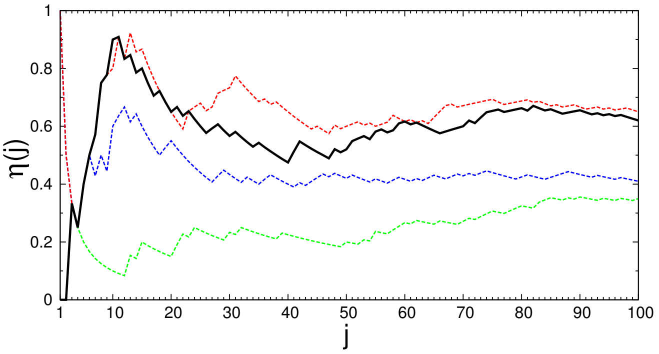

Fig. 1. Overlap $\eta(j) = j_c/j$ of WRWU with ARWU as a function of rank index j of ARWU. Here index j_c is the number of common universities in the top j indexes of two rankings; curves show the overlap between: WPRWU and ARWU (black curve), ARWU and top PageRank universities of English, French, German (red dashed, blue dashed, green dashed curves, respectively) Wikipedia editions. The overlaps with ARWU for WPRWU of English and French Wikipedia edition ranks are superimposed up to $j = 6$ (black, red and blue dashed curves), and for ARWU and WPRWU of German edition the ranks are superimposed up to $j = 4$ (black and green dashed curves).

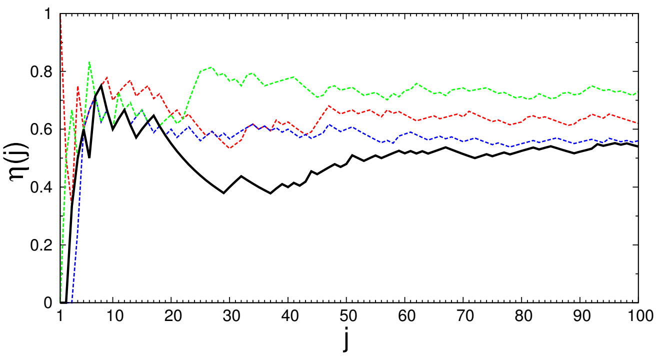

Fig. 2. Overlap $\eta(j) = j_c/j$ between: W2RWU and ARWU (black curve), W2RWU and WPRWU (red dashed curve), W2RWU and WCRWU (blue dashed curve), WPRWU and WCRWU (green dashed curve). Here j is the rank index of both compared ranks, j_c is number of common items among j ranks.

Oxford and Harvard which have rank positions $K_{ARWU} = 5, 10, 1$ respectively. We will see later that WRWU gives more favor to ancient universities, comparing to ARWU. The ranks WCRWU and W2RWU have smaller overlap with ARWU. It is related to the fact that these ranks incorporate the communication features of the article since CheiRank highlights the effect of outgoing links. Thus we see that certain university (their articles) are not very communicative (relatively small number of important outgoing links; e.g. Chicago, Berkeley) so that they do not enter in top 10 of WCRWU and W2RWU. In contrast we see that there are a few non-Anglo-Saxon universities which gain their higher positions in W2RWU and WCRWU being more communicative.

The dependence of overlap fractions $\eta(j)$ between two university ranks up to index j is shown in Fig. 1 for ARWU and WPRWU, ARWU and PageRank of English, French, and German editions. For ARWU and WPRWU we find $\eta(100) = 0.62$. It is interesting to note that English edition has a larger overlap with ARWU ($\eta(100) = 0.65$), followed by French ($\eta(100) = 0.41$) and German ($\eta(100) = 0.35$) editions. Thus we see that ARWU highlights in a stronger way the contribution of EN universities while FR and DE editions highlight in a stronger way the universities of their languages. Indeed, we have in the top 100 lists 32 French and 63 German speaking universities for FR and DE editions (see [27]). This demonstrates significantly different cultural views developed in each language edition. We will see below that there are also strong historical reasons behind such cultural preferences.

The overlap between ARWU and W2RWU of 2DRank is shown in Fig. 2. There is a notable reduction of η comparing to the case of WPRWU of Fig. 1 well visible at $j \approx 10$ and $j = 100$ where we have $\eta(10) = 0.6$, $\eta(100) = 0.54$ instead of larger WPRWU values given above. The overlaps between W2RWU and WPRWU and WCRWU lists have larger values due to certain correlations between PageRank and CheiRank vectors discussed for Wikipedia networks at [14,18,29].

All 100 universities from 24 editions are ordered by their respective indexes of WPRWU PageRank K_U and WCRWU CheiRank K^*_U. Their distribution on the PageRank - CheiRank plane (K_U, K^*_U) is shown in Fig. 3 for the top 1000 universities. In contrast to a very broad distribution of all Wikipedia articles on this plane (see Figs. in [14, 18,21]), for universities we have significantly stronger concentration around diagonal ($K_U = K^*{}_U$). It looks like that ingoing and outgoing information for articles of universities is approximately conserved like it is approximately the case of commercial flows on the world trade network where the countries try to keep their economic balance [30]. Thus we can say that universities of Oxford, Yale, Uppsala are more communicative (located below diagonal) while those of Cambridge, Princeton, Chicago are much less communicative (located above diagonal). This presentation shows that certain universities have open possibilities for improvement of communicative flows of their Wikipedia articles.

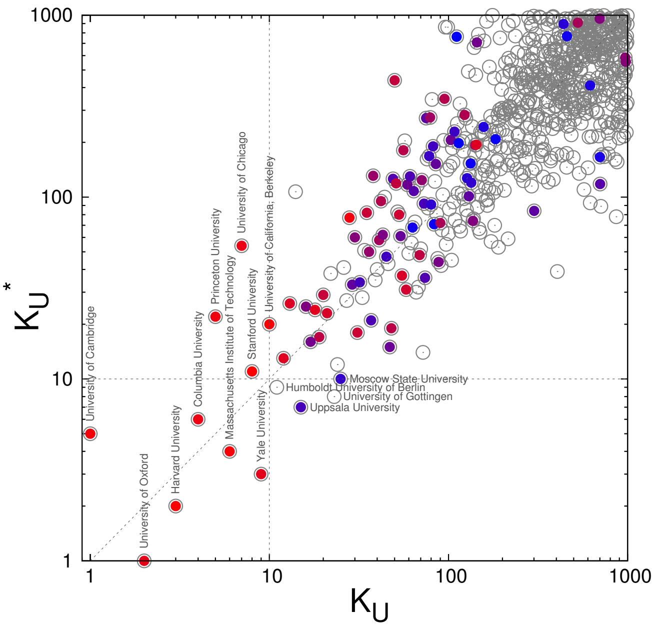

Fig. 3. Distribution of world universities on the PageRank - CheiRank plane (K_U, K_U^*) (open circles), where K_U, K_U^* are ranks of a given University U in WPRWU, WCRWU. Universities appearing in top 100 ARWU Shanghai ranking are shown by colored full circles with the color ranging from red (ARWU rank 1) to blue (ARWU rank 100). The names of certain universities are given.

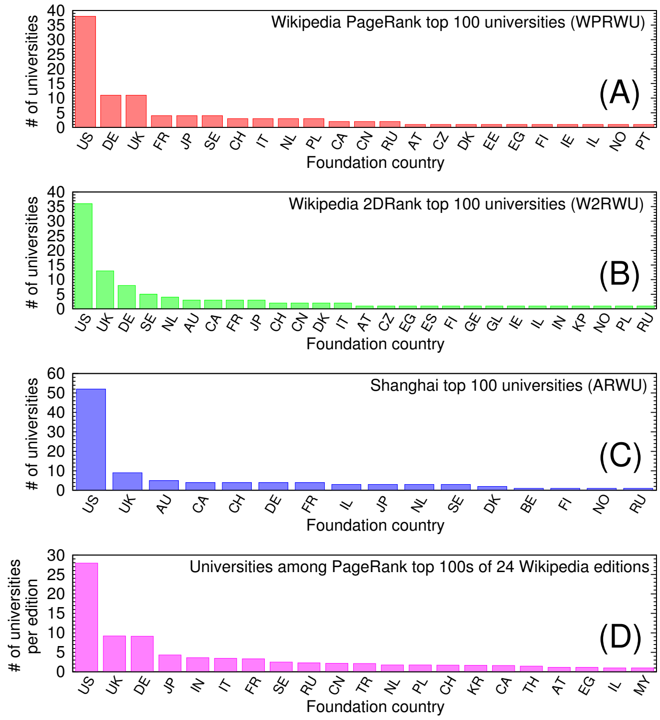

Fig. 4. Distribution over foundation countries of top 100 universities from WPRWU (A), from W2RWU (B) and from ARWU (Shanghai ranking) (C). Panel (D) shows the average (per edition) number of universities founded in a given country appearing in the PageRank top 100 universities of 24 Wikipedia editions. In panel (D) the countries with the score less than 1 are not represented.

The distribution of ARWU top 100 universities is also displayed on (K_U, K_U^*) plane in Fig. 3. It shows that ARWU universities are located mainly at top K_U, K_U^* indexes. However, some universities, located at top K_U, K_U^* positions, are absent in ARWU list. For example, these are Humboldt and Göttingen Universities. Such cases stress the important difference between ARWU and WRWU. Namely, WRWU gives credit to historically important universities which played an important role during the whole human history (e.g. the two cases above are definitely important for German and EU history) while ARWU gives much more importance to instantaneous achievements. We return to discussion of such differences in next Sections.

Finally, we note that the W2RWU list is not composed simply from the items of top squares of (K_U, K_U^*) plane, as in the usual 2DRank algorithm [18], because W2RWU is obtained via the relation (4) which performs averaging over 24 editions.

3.2 Geographical distribution

According to Table 2 we attribute to each country a corresponding language edition choosing mostly used language in a country defined by actual country borders (see [28]). We determine a foundation country, with its actual borders, at which a university has been founded. Then from top 100 universities of WPRWU, W2RWU, ARWU we obtain their distributions over the world countries shown in Fig. 4. The top 3 countries are US, DE, UK and US, UK, DE for WPRWU and W2RWU respectively. This is rather different from the top 3 countries US, UK, AU of ARWU. Also the weight of US is significantly reduced from 52 percent for ARWU to 38 and 36 percent for WPRWU and W2RWU respectively (see Fig. 4A,B,C). If we consider the average over all 24 editions then we get on the top 3 positions US, UK, DE with even smaller 28.2 percent of US (see Fig. 4D). Thus Wikipedia ranking provides a more balanced cultural view on important universities. Indeed, each edition gives more “votes” for universities of same language that increases contributions of various languages (or cultures) even if other cultures do not necessarily agree on importance of such a choice, thus introducing certain contrebalance.

The distribution of all top 100 PageRank universities of 24 editions over the world map of countries [28] is shown in Fig. 5. The similar world map for WCRWU is shown in Fig. 6. For WCRWU we see appearance of new countries in Africa and Central Asia being related to a larger number of outgoing links of their universities. Differently from top 100 of WPRWU (or WCRWU), here all 24 editions give a more significant contribution with a noticeable weight for India and Japan. Indeed, in HI and JA edition rankings there are large fractions of universities of their own languages (81 and 65 percent respectively) that leads to their weight increase among all 1025 PageRank universities. However, this effect of self citations is signifi-

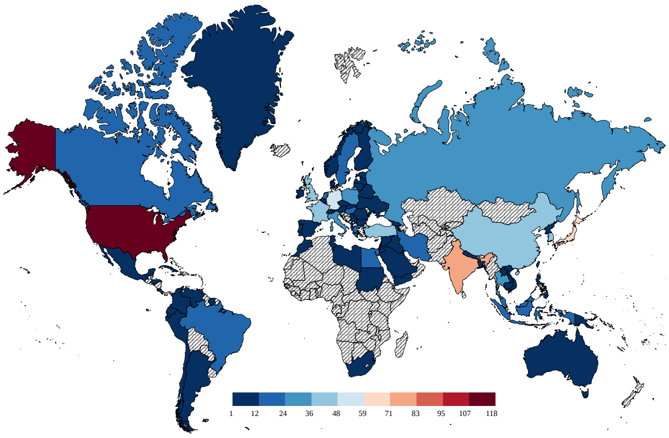

Fig. 5. Geographical distribution of universities appearing in the top 100 universities of all 24 Wikipedia editions given by PageRank algorithm. The total number of universities is 1025. Colors range from dark blue (small number of universities) to dark red (maximum number of universities, here 118 for US). Countries filled by dashed lines pattern have no university in the top 100 lists of 24 editions.

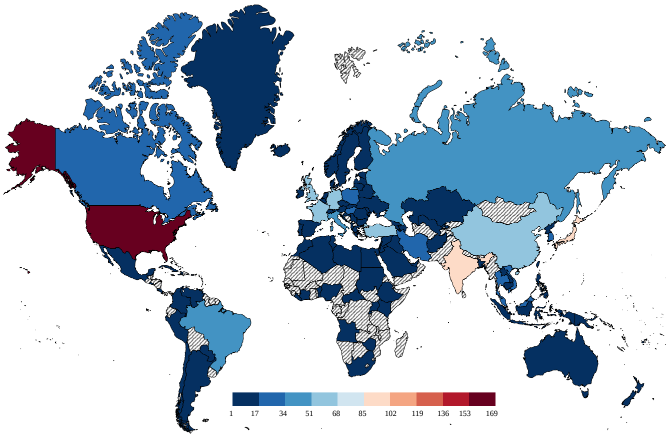

Fig. 6. Geographical distribution of universities appearing in the top 100 universities of all 24 Wikipedia editions given by CheiRank algorithm. The total number of universities is 1379. Colors range from dark blue (small number of universities) to dark red (maximum number of universities, here 169 for US). Countries filled by dashed lines pattern have no university in the top 100 lists of 24 editions.

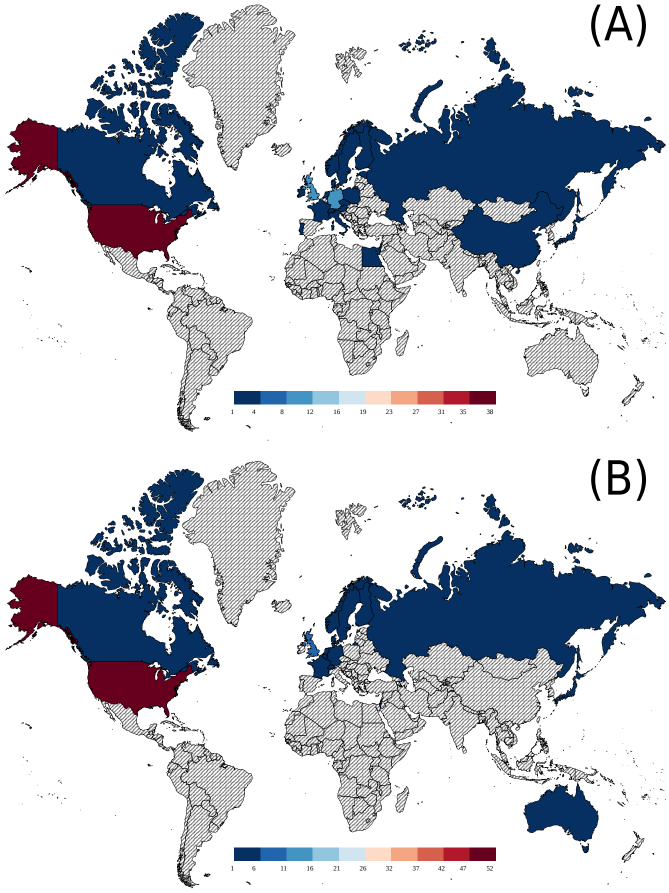

Fig. 7. Geographical distribution of top 100 universities of WPRWU from PageRank algorithm (A) and of ARWU Shanghai ranking (B). Colors range from dark blue (small number of universities) to dark red (maximum 38 (A) and 52 (B) for US). Countries filled by dashed lines pattern have no university in corresponding top 100.

cantly suppressed for top 100 of WPRWU where opinions of other editions play a role. Thus Indian universities do not appear in the WPRWU top 100 as it is seen in Fig. 4 and in the corresponding world map of Fig. 7(A) showing WPRWU top 100 universities.

The geographical distributions of WPRWU and ARWU are shown in Fig. 7 (A) and (B). We see that Australia, present at high positions in ARWU, is not present at WPRWU, while inversely China is present on WPRWU map being absent in ARWU. Also an absolute percent of US is significantly reduced in WPRWU with Germany taking the second position at WPRWU instead of 6th position in ARWU.

Another interesting comparison of efficiency of universities is given by a number of top 100 universities per 10 millions inhabitants for a given country (the actual country population is taken mainly in 2015 from [31], see also [27]). These distributions for highly raked countries are shown in Fig. 8 for WPRWU and ARWU. At the top 3 positions we find Estonia, Sweden, Switzerland for WPRWU and Switzerland, Israel, Denmark for ARWU. Estonia appears on the top due to its small population and the only one University of Tartu. This ancient university, founded in 1632, was historically on the crossroads of Sweden, Russia, Poland, Germany thus being important for various cultures in this region. Now it is located in Estonia with its small population that explains its top WPRWU per inhabitant position. The example of University of Tartu highlights the importance of historical environment for appreciation of role of a given university. It shows that WPRWU takes into account the history of university while ARWU ignores this feature.

On the second WPRWU position per inhabitant we have Sweden with 4 universities in top 100. Sweden is followed by Switzerland with 3 universities. For ARWU

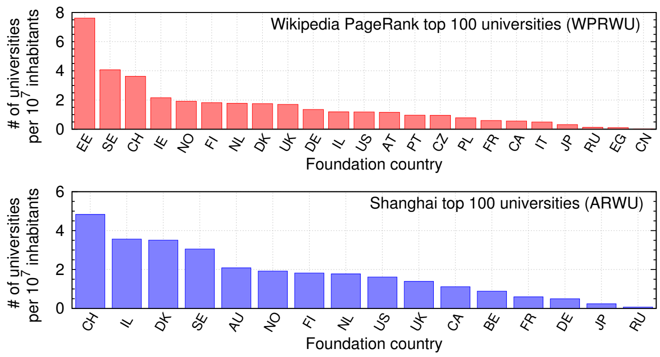

Fig. 8. Distributions of WPRWU and ARWU top 100 universities per number of inhabitants (measured in 10 millions) over foundation country.

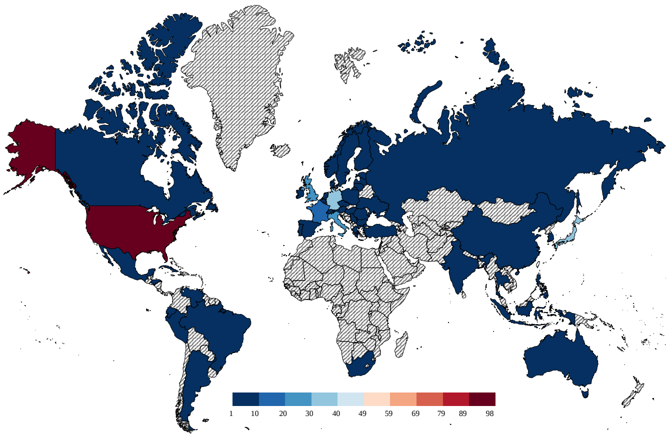

Fig. 9. Geographical distribution of universities appearing in the top 100 universities of all 24 Wikipedia editions given by PageRank algorithm and founded before 20th century (384 in total). Colors range from dark blue (small number of universities) to dark red (maximum number of universities, here 98 for US). Countries filled by dashed lines pattern have no university in the top 100 lists of 24 editions.

Switzerland is at the top position with 4 universities. We see that the Northern countries are taking high positions (SE, IE, NO, FI ...) both for WPRWU and ARWU.

We now go to analysis of time evolution of top universities.

3.3 Evolution through centuries

Each university has its year of foundation and thus we attribute all universities to their own foundation century. From top 100 universities of all editions with PageRank we select universities founded before 20th century (384 in total) and present their geographical distribution over world countries in Fig. 9. The comparison with Fig. 5 shows a significant drop of number of universities in Africa (only South Africa remains), India and Japan lose their high positions while EU countries (DE, UK, IT, FR) are improving their positions but US still takes the first top position with the largest number of universities. The geographical distribution for universities founded before 19th century (139 in total) is shown in Fig. 10. Here Germany is taking the top position followed by Italy. But already US takes the 3rd position.

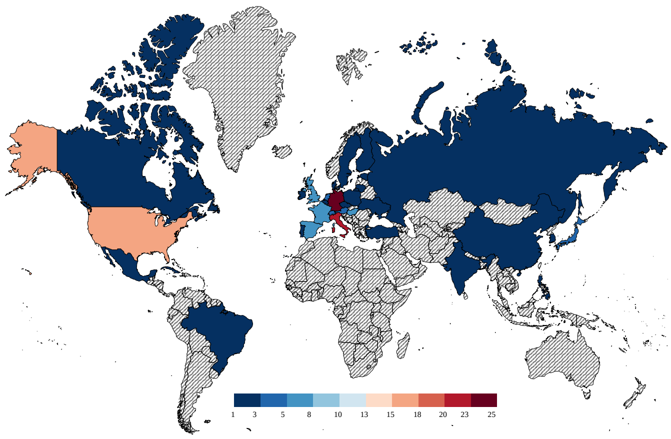

Fig. 10. Same as in Fig. 9 but with the foundation date before 19th century (139 universities in total; maximum is for DE with 25 universities).

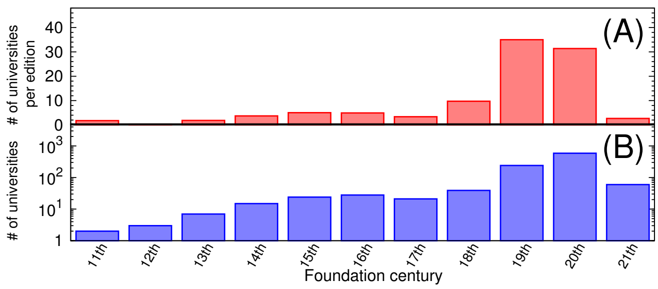

Fig. 11. Distribution over foundation century for universities N_f, appearing in the PageRank top 100 universities of 24 Wikipedia editions (1025 in total); panel (A) gives the average number per edition N_{fe}, panel (B) gives the total number of universities N_f founded in a given century.

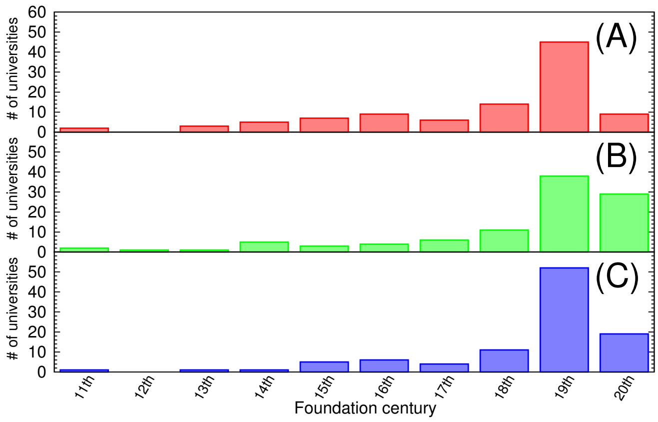

Fig. 12. Distribution over foundation century of top 100 universities from WPRWU (A), W2RWU (B) and ARWU (C).

The distribution of number of universities N_f founded in a given century is given in Fig. 11. It shows that both, the total number N_f and average number of universities

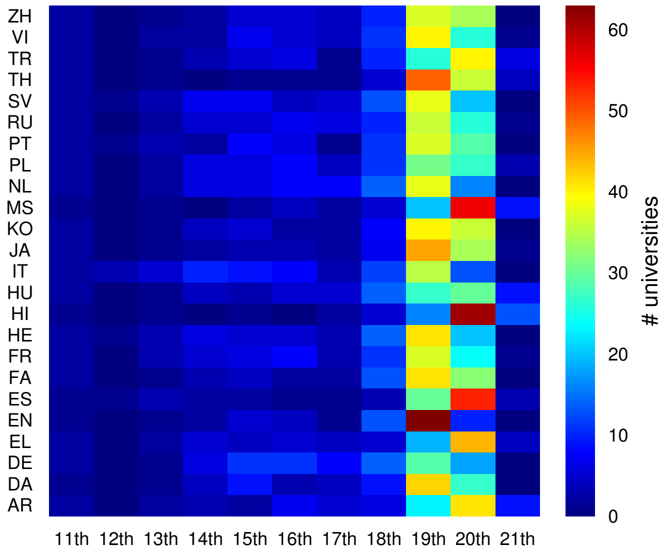

Fig. 13. Distribution over foundation century for PageRank top 100 universities of each Wikipedia edition. Dark red color corresponds to maximum with 63 for EN in 19th century.

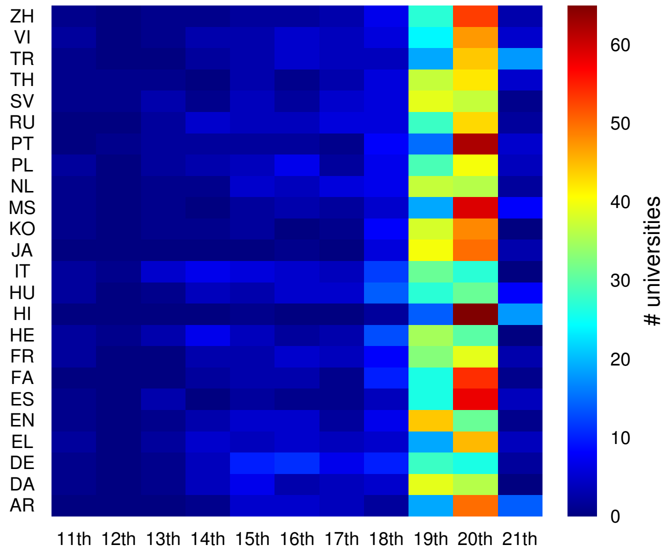

Fig. 14. Distribution over foundation century for CheiRank top 100 universities of each Wikipedia edition. Dark red color corresponds to maximum with 65 for HI in 20th century.

per edition N_{fe}, remain approximately constant during 14th to 17th centuries. The steady growth of N_f starts from 18th century being close to an exponential increase. On the other side if we consider only top 100 of global ranking of WPRWU, W2RWU and ARWU, shown in Fig. 12, then we see that the main part of top universities has been founded in 19th century (about 50) and in 20th century there appeared only about 10 to 20 universities which succeed to enter in the top 100 list. Thus we see that the top 100 club is rather rigid in accepting new "members" with time. Only for W2RWU there is some redistribution in 20th century mainly because the new young universities are more communicative that improves their 2DRank positions. In contrast for WPRWU we have 43 (5) first universities, founded before 20th (19th) century, which remained at unchanged positions now. In total in WPRWU list there are 8 and 54 universities founded in 20th, and 19th or earlier century. This confirm the highly rigid nature of top 100 positing of leading universities which was mainly formed before 20th century.

In Fig. 12 we consider the global list of top 100 universities. The data for top 100 of each edition at each foundation century are presented in Fig. 13 for PageRank and in Fig. 14 for CheiRank. The data of Fig. 13 show emergence of many universities in PageRank top 100 in 20th century for HI, MS and ES. In contrast, there are practically no new universities in 20th century for EN, IT, NL showing that their contribution to the top 100 list is dominated by 19th century. The situation is different for the CheiRank top 100 in Fig. 14: here there are many universities appearing in 20th century especially for HI, PT, MS and even ES. We attribute this to the fact that CheiRank highlights the communicative features of Wikipedia articles and that the new young universities are more better placed in communicative broadcast activity.

4 Entanglement and interactions of cultures

The results for Wikipedia ranking in different editions can be used for analysis of entanglement and interactions of cultures. We associate each language to a culture since it represents the most important cultural feature. Such an approach was used in [21,22] for historical figures and this method can be also directly used for universities. For that we count how many universities N_{ij}, attributed to a culture (language) i, appears in the top 100 universities of culture (language) j. This gives the number of directed links from node j to node i and then the Google matrix of cultures is constructed in the usual way (1) with the same damping factor $\alpha = 0.85$ (see also [21,22]). In total we have 25 nodes, since some universities cannot be attributed to any of 24 editions and corresponding cultures and in such a case we attribute them to an additional culture WR. The diagonal self citations inside the same culture are not considered so that $N_{ii} = 0, 1 \leq i \leq 25$. The number N_{ij} can be defined for any of the three ranking algorithms discussed above.

The network of cultures, constructed from top 100 universities of PageRank, is shown in Fig. 15. For the CheiRank algorithm the network of cultures is shown in Fig. 16. For the presentation of these directed networks we use gephy software with a circular layout [32]. These network images show strong interconnections between different cultures.

To analyze this complex entanglement of cultures we determine the PageRank and CheiRank vectors of such a secondary network with 25 nodes. The size of a node in Fig. 15 (Fig. 16) is proportional to PageRank (CheiRank) probability. In addition we determine PageRank K and CheiRank K^* indexes of cultures and display all $24+1$ cultures on the (K, K^*) plane in Fig. 17. For PageRank list (Fig. 17A) we have the strong dominance of

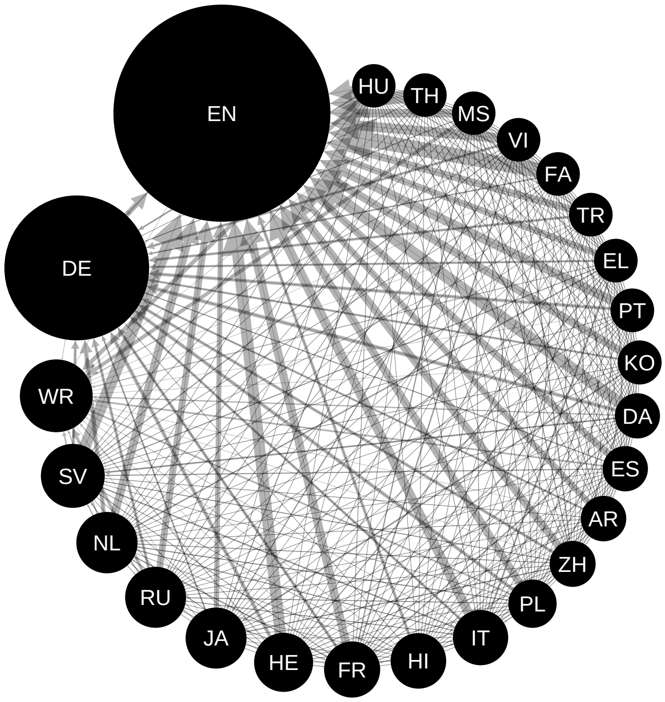

Fig. 15. Network of cultures constructed from the PageRank top 100 universities of 24 Wikipedia editions. The width of each directed link is proportional to the number of foreign universities quoted in top 100 of a given culture; links inside cultures are not considered. The size of a node is proportional to its PageRank value.

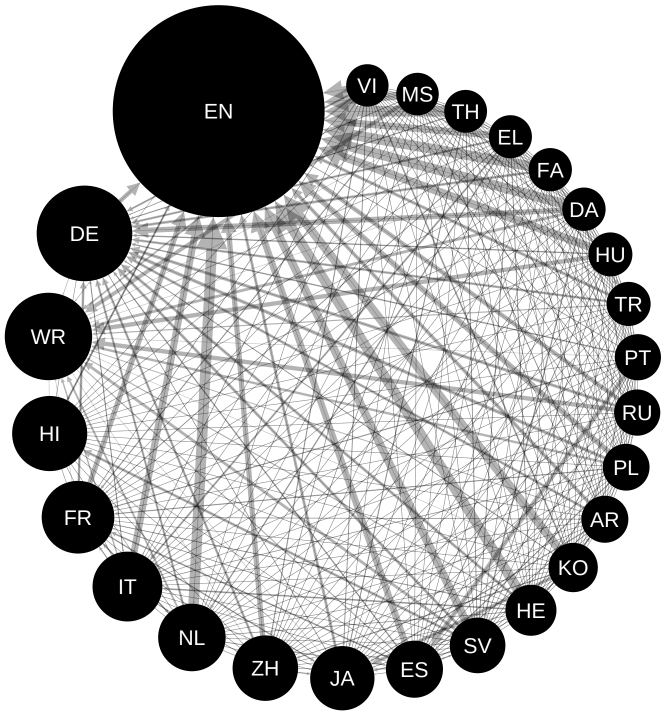

Fig. 16. Same as in Fig. 15 but for the CheiRank top 100 universities of 24 Wikipedia editions.

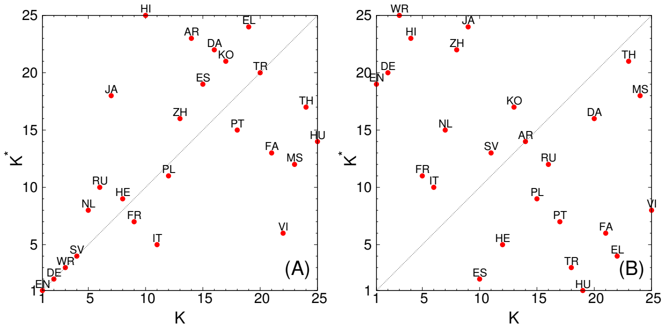

Fig. 17. PageRank-CheiRank (K, K^*) plane for the networks of cultures from PageRank universities of Fig. 15 (panel A) and from CheiRank universities of Fig. 16 (panel B). Each culture is marked by language code.

EN, DE, WR, and SV cultures which take the top 4 positions being on the diagonal $K = K^*$. Next positions in K are taken by NL, RU and in K^* by IT, FR. At highest positions of K and K^* we have HI and HU which have many self citations. For CheiRank list (Fig. 17B) we have significantly broader distributions of cultures on (K, K^*) plane with EN, DE in top K positions and HU, ES in top K^* positions. We assume that many well-know scientists immigrated from Hungary, and many Spanish speaking countries in Latin America are responsible for stronger communicative features of HU and ES while EN and DE still keep their top PageRank positions. We note that for ranking of universities we obtain the distributions of cultures over K, K^* plane being rather different from the distributions of cultures obtained from ranking of historical figures (see Fig.10 in [22]). This shows that the entanglement of cultures takes place on various levels of knowledge having complex interactions on each level. Thus appreciation of foreign universities in a given culture works in a rather different manner comparing to appreciation of foreign historical figures.

5 Discussion

In this work we presented the Wikipedia ranking of world universities using PageRank, 2DRank and CheiRank algorithms developed for directed networks where they proved their efficiency. The analysis is based on 24 Wikipedia language editions that allows to take into account various cultural view points. At the same time these cultural views are considered by the statistical mathematical analysis of all human knowledge accumulated in these 24 editions containing 17 millions Wikipedia articles. Thus our analysis gives no cultural preferences standing on pure mathematical grounds.

We find that the PageRank list of WPRWU top 100 universities has 62 percent overlap with ARWU Shanghai list demonstrating that this analysis gives reliable results. At the same time WPRWU gives more emphasis to non-Anglo-Saxon cultures reducing the percent of US universities from 52 in ARWU to 38 in WPRWU. Our results show that German universities take the 2nd position in this ranking with UK, France, Japan and Sweden taking next 3rd to 6th places, while ARWU gives respectively the places 6th for DE, 2nd for UK, 7th for FR, 9th for JP and 11th for SE. The number of top PageRank universities per inhabitant demonstrates the efficiency of universities in countries of Northern Europe and Switzerland.

The rankings based on 2DRank and CheiRank algorithms highlight in a better manner the communicative and broadcast features of universities showing that their efficiency varies strongly even for top ranked universities.

The analysis of university ranking evolution through ten centuries shows that Wikipedia highlights significantly stronger historically important universities which role is reduced in ARWU. Thus for PageRank list of top 100 universities in 24 editions we find the dominance of Germany and Italy before 19th century, even if the rise of US universities is already visible to that times. The dominance of US is established after 19th century. Our WRWU results show that the club of top universities is formed mainly before 20th century and that it remains very ridig in "accepting" new members after that time.

The appreciation of foreign universities in individual editions allows to determine effective interactions of 24 cultures related to language editions showing the strong influence of English, German and Swedish universities.

We think that the Wikipedia ranking provides the firm mathematical statistical evaluation of world universities which can be viewed as a new independent ranking being complementary to already existing approaches. In the view of importance of university ranking for higher education [2] we hope that the WRWU method will also find a broad usage together with other rankings.

References

1. P.G. Altbach, L. Reisberg and L.E. Rumbley, *Trends in Global Higher Education: Tracking an Academic Revolution. A Report Prepared for the UNESCO 2009 World Conference on Higher Education*, UNESCO (2009), Available: `http://www.unesco.org/fileadmin/MULTIMEDIA/HQ/ED/ED/pdf/WCHE_2009/`. Accessed November 2015
2. E. Hazelkorn, *Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence*, Palgrave Macmillan, New York (2015)
3. Academic Ranking of World Universities. Available: `http://www.shanghairanking.com/`. Accessed November 2015
4. Enseignement supérieur et recherche, Investissements d'avenir. Available: `http://www.enseignementsup-recherche.gouv.fr/pid24578/investissements-d-avenir.html`. Accessed November 2015
5. Russian Academic Excellence Project. Available: `http://5top100.ru/`. Accessed November 2015
6. Times Higher Education World University Ranking. Available: `https://www.timeshighereducation.com`. Accessed November 2015
7. U-Multirank of European Union. Available: `http://www.umultirank.org/`. Accessed November 2015
8. IREG Observatory on Academic Ranking and Excellence. Available: `http://ireg-observatory.org/en/`. Accessed November 2015
9. D. Docampo, Scientometrics **86**, 77 (2011)
10. L. Bornmann, R. Mutz and H.-D. Daniel, J. American Society Information Science and Technology, **64**, 1649 (2013)
11. H. Jöns and M. Hoyler, Geoforum **46**, 45 (2013)
12. D. Docampo and L. Cram, Scientometrics **98**, 1347 (2014)
13. A. Rauhvargers, *Global University Rankings and Their Impact - Report II*, European University Association asbl (2013) (2013). Available: `http://www.eua.be/`. Accessed November 2015.
14. A.O. Zhirov, O.V. Zhirov and D.L. Shepelyansky, Eur. Phys. J. B **77**, 523 (2010).
15. A.A. Markov, *Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga*, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, **15**, 135 (1906) (in Russian) [English trans.: *Extension of the limit theorems of probability theory to a sum of variables connected in a chain* reprinted in Appendix B of: R.A. Howard *Dynamic Probabilistic Systems*, volume 1: *Markov models*, Dover Publ. (2007)]
16. S. Brin and L. Page, Computer Networks and ISDN Systems **30**, 107 (1998).
17. A.M. Langville and C.D. Meyer, *Google's PageRank and beyond: the science of search engine rankings*, Princeton University Press, Princeton (2006)
18. L. Ermann, K.M. Frahm and D.L. Shepelyansky, Rev. Mod. Phys. **87**, 1261 (2015)
19. S. Dorogovtsev, *Lectures on complex networks*, Oxford University Press, Oxford (2010).
20. Y.-H. Eom, K.M. Frahm, A. Benczur and D.L. Shepelyansky, Eur. Phys. J. B **86**, 492 (2013)
21. Y.-H. Eom and D.L. Shepelyansky, PLoS ONE **8(10)**, e74554 (2013)
22. Y.-H. Eom, P.Aragon, D.Laniado, A.Kaltenbrunner, S.Vigna and D.L. Shepelyansky, PLoS ONE **10(3)**, e0114825 (2015)
23. M.H. Hart, *The 100: ranking of the most influential persons in history*, Citadel Press, N.Y. (1992)
24. D.L.Shepelyansky, *Top 100 historical figures of Wikipedia*, Available: `https://hal.archives-ouvertes.fr/hal-01184245/` (2015) Accessed November 2015.
25. Country codes ISO 3166-1 alpha-2. Available: `https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2`. Accessed November 2015
26. A.D. Chepelianskii, *Towards physical laws for software architecture*, arXiv:1003.5455 [cs.SE] (2010)
27. WRWU rank data. Available: `http://perso.utinam.cnrs.fr/~lages/datasets/WRWU/`. Accessed November 2015.
28. Web page *Maps of the world* Available: `http://www.mapsofworld.com`. Accessed November 2015.
29. L. Ermann, A.D. Chepelianskii and D.L. Shepelyansky, J. Phys. A: Math. Theor. **45**, 275101 (2012)
30. L. Ermann and D.L. Shepelyansky, Acta Physica Polonica A **120(6A)**, A158 (2011)
31. Wikipedia, The Free Encyclopedia, *List of countries and dependencies by population.* (2015). Available `https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population`. Accessed November 2015.
32. M. Bastian, S. Heymann, M. Jacomy, *Gephi: an open source software for exploring and manipulating networks*, International AAAI Conference on Weblogs and Social Media (2009)