Reduced Google Matrix analysis of Wikipedia networks

Katia Jaffrès-Runser

In collaboration with Samer El Zant, Dima Shepelyanski and Klaus Frahm.

IRIT, INP-ENSEEIHT, University of Toulouse

Conference
Google Matrix: fundamentals, applications and beyond
Institut des Hautes Études Scientifiques
Le Bois-Marie, 16 octobre 2018

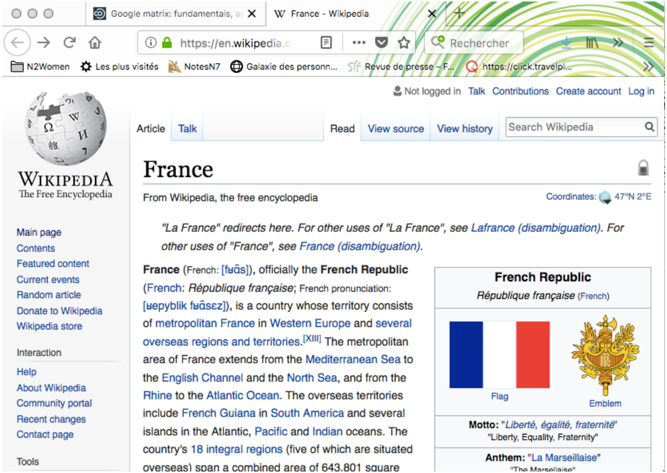
IRIT Institut de Recherche en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

INP TOULOUSE ENSEEIHT

Université Fédérale
Toulouse Midi-Pyrénées

Wikipedia

Is a free, collaboratively written encyclopædia



France

From Wikipedia, the free encyclopedia

Coordinates: 47°N 2°E

"La France" redirects here. For other uses of "La France", see Lafrance (disambiguation). For other uses of "France", see France (disambiguation).

France (French: [fʁɑ̃s]), officially the **French Republic** (French: *République française*; French pronunciation: [ʁepyblik fʁɑ̃sɛz]), is a country whose territory consists of metropolitan France in Western Europe and several overseas regions and territories.[XIII] The metropolitan area of France extends from the Mediterranean Sea to the English Channel and the North Sea, and from the Rhine to the Atlantic Ocean. The overseas territories include French Guiana in South America and several islands in the Atlantic, Pacific and Indian oceans. The country's 18 integral regions (five of which are situated overseas) span a combined area of 643,801 square

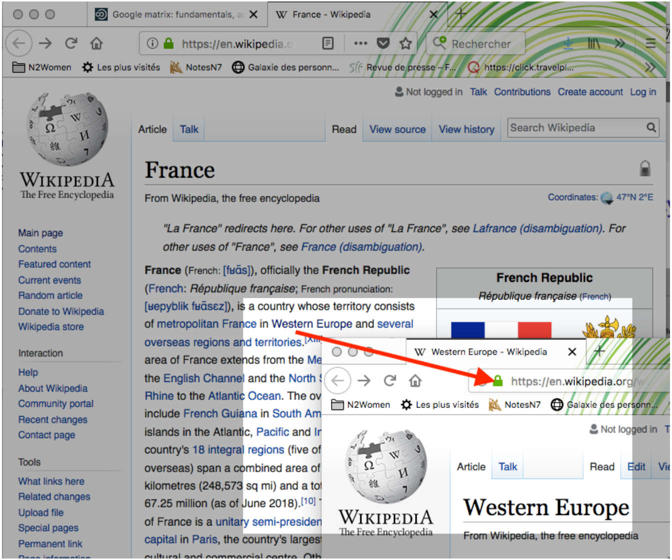
French Republic
République française (French)

Motto: "*Liberté, égalité, fraternité*"
"Liberty, Equality, Fraternity"

Anthem: "La Marseillaise"
"The Marseillaise"

Wikipedia

Offers a hyperlinked structure for all articles



France

From Wikipedia, the free encyclopedia

Coordinates: 47°N 2°E

"La France" redirects here. For other uses of "La France", see Lafrance (disambiguation). For other uses of "France", see France (disambiguation).

France (French: [fʁɑ̃s]), officially the **French Republic** (French: *République française*; French pronunciation: [ʁepyblik fʁɑ̃sɛz]), is a country whose territory consists of metropolitan France in Western Europe and several overseas regions and territories.[XI] area of France extends from the Me the English Channel and the North Rhine to the Atlantic Ocean. The ov include French Guiana in South Am islands in the Atlantic, Pacific and In country's 18 integral regions (five of overseas) span a combined area of kilometres (248,573 sq mi) and a to 67.25 million (as of June 2018).[10] of France is a unitary semi-presiden capital in Paris, the country's larges cultural and commercial centre. Oth

French Republic

République française (French)

Western Europe

From Wikipedia, the free encyclopedia

Wikipedia

That can be directly mapped to a directed network of topics that is scale-free.

Wikipedia edition	Number of nodes	Number of links
Arabic 2013	203 326	1 896 621
English 2013	4 212 493	101 611 731
English **2017**	5 416 537	122 232 932
French 2013	1 352 825	34 431 943
German 2013	1 532 977	36 781 077
Italic 2013	1 017 953	25 667 781
Russian 2013	966 284	20 853 206
Spanish 2013	974 021	23 105 758

Table: **Wikipedia editions and their sizes.**

Google matrix analysis of Wikipedia

Google matrix

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

S is the matrix of Markov transitions with $S_{ij} = A_{ij}/k_{out}(j)$ giving the probability of moving from article j to i.
A is the adjacency matrix and k_{out} the out-degree.
If j is a dangling node : $S_{ij} = 1/N$.

PageRank eigenvector

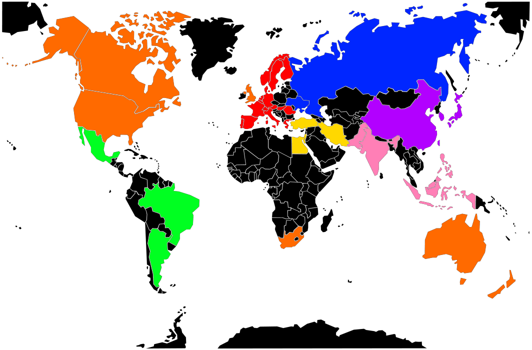
Captures **central nodes** in Wikipedia

- Ranking of historical figures over 35 centuries ($\sim$ Hart ranking)
- Ranking of world universities ($\sim$ Shanghai Academic ranking)

Good **diffusion nodes** are identified with CheiRank
(using G^* derived from the transposed version of A).

PageRank example

Top 40 countries in PageRank for EnWiki



PR Index	1	2	3	4	5	6	7	8	9	10	11	12	...
CC	US	FR	GB	DE	CA	IN	AU	IT	JP	CN	RU	ES	...

Accounting for different editions

PageRank values depend on editions of Wikipedia:

PR Index	1	2	3	4	5	6	7	8	9	10	...
EnWiki	US	FR	GB	DE	CA	IN	AU	IT	JP	CN	...
RuWiki	RU	US	FR	DE	UA	IT	GB	ES	CN	PL	...

Cross-edition rank: $\Theta-$score

Have a global ranking across several editions.

$$\Theta_P = \sum_E (101 - R_{P,E}). \tag{1}$$

Here $R_{P,E}$ is the ranking of top 100 nodes in edition E of Wikipedia. The largest Θ_P, the most important the node is accros all editions.

Accounting for different editions

Top 40 painters

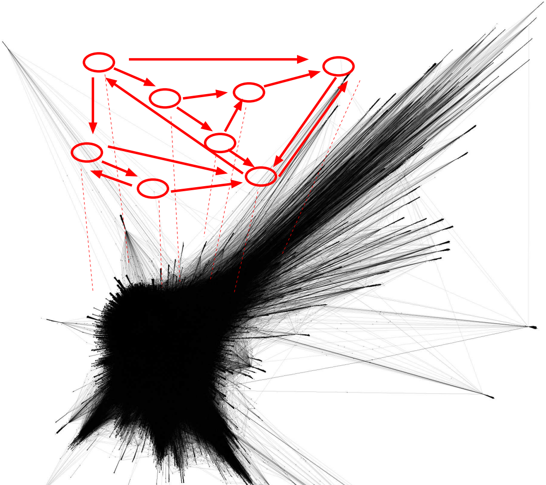
$\Theta-$score over 7 editions: EnWiki, FrWiki, RuWiki, DeWiki, ItWiki, EsWiki and NlWiki.

Θ rank	K_{av} rank	Painter	Θ rank	K_{av} rank	Painter
1	1	Vinci	21	18	Bondone
2	2	Picasso	22	25	Kandinsky
3	6	Gogh	23	19	Botticelli
4	4	Rembrandt	24	21	Caravaggio
5	5	Rubens	25	23	Velázquez
6	8	Durer	26	30	Degas
7	9	Titian	27	26	Bruegel Eld
8	11	Monet	28	29	Dyck
9	12	Dali	29	28	Renoir
10	14	Cézanne	30	31	Chagall
11	3	Michelangelo	31	33	Lautrec
12	7	Raphael	32	27	Vermeer
13	10	Goya	33	36	Poussin
14	13	Vasari	34	37	Turner
15	16	Matisse	35	38	Braque
16	15	Warhol	36	32	Blake
17	17	Delacroix	37	34	Greco
18	22	Manet	38	39	Miró
19	20	David	39	35	Munch
20	24	Gauguin	40	40	Eyck

Google matrix analysis of Wikipedia

Reduced Google matrix

A powerful tool to create a sub-network (or *thematic view*) of the full Google matrix for a set of N_r articles.



Google matrix analysis of Wikipedia

Reduced network

Network decomposition into

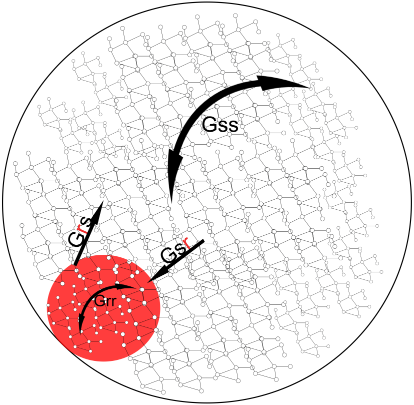
- Reduced network of N_r nodes
- Rest of nodes $N_s = N - N_r$

Reordering G, we have :

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

And corresponding PageRank:

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



Google matrix analysis of Wikipedia

Reduced network

Network decomposition into

- Reduced network of N_r nodes
- Rest of nodes $N_s = N - N_r$

Reordering G, we have:

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

And corresponding PageRank:

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

Reduced Google matrix G_{R}

We want:

$$G_{\mathrm{R}} P_r = P_r$$

And thus, if G_{ss} is not singular, we have:

$$G_{\mathrm{R}} = G_{rr} + G_{rs}(\mathbf{1} - G_{ss})^{-1} G_{sr}$$

Numerical evaluation of G_{R}

N_s is too large for a direct evaluation of $(1 - G_{ss})^{-1}$.

The following numerical evaluation has been proposed by Klaus Frahm:

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

where λ_c is the leading eigenvalue of G_{ss}, $\mathcal{P}_c$ the projector onto the eigenspace of λ_c and $\mathcal{Q}_c$ the complementary projector.

Components of G_{R}

G_{R} has 3 components

$$G_{\mathrm{R}} = G_{rr} + G_{\mathrm{pr}} + G_{\mathrm{qr}},$$

with:

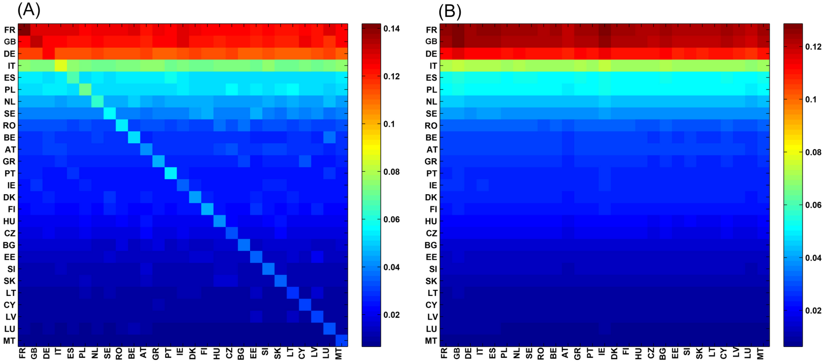
- G_{rr} the **direct** interactions within the sub-network
- $G_{\mathrm{pr}} = G_{rs}\mathcal{P}_c G_{sr}/(1-\lambda_c)$, the **projector** component
- $G_{\mathrm{qr}} = G_{rs}[\mathcal{Q}_c \sum_{l=0}^{\infty} \bar{G}_{ss}^{\,l}]G_{sr}$, the **indirect** interactions through the rest of nodes.

27 EU countries for EnWiki

$$\mathbf{G}_{\mathrm{R}} = G_{rr} + \mathbf{G}_{\mathrm{pr}} + G_{\mathrm{qr}}$$

Reduced Matrix G_{R}

Projector component G_{pr}



Column sums of G_{pr} account for $\sim$ 95-97% of the total column sum of G_{R}.

27 EU countries for EnWiki

Total weight of matrices

Sum of all elements of corresponding matrices for the 27 EU network and the 40 top worldwide set of countries.

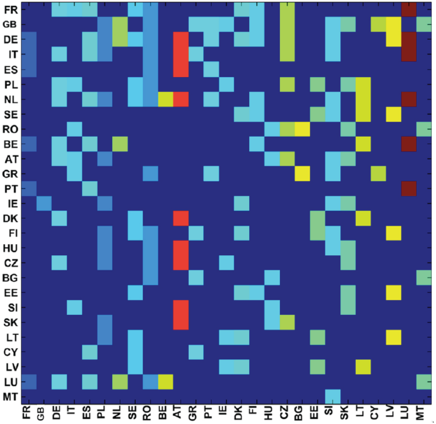
	$W_{\rm pr}$	$W_{\rm qr}$	W_{rr}	Sum
40 worldwide	0.96120	0.029702	0.009098	1
27 EU	0.95332	0.038346	0.008334	1

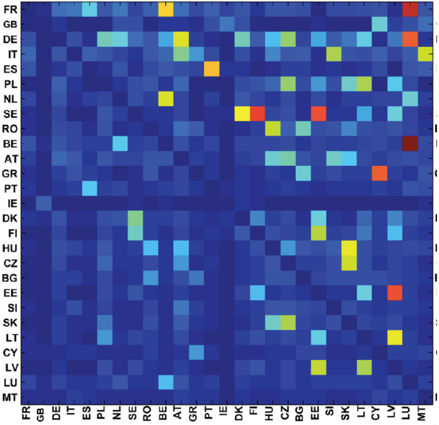
27 EU countries for EnWiki

$$G_{\mathrm{R}} = \mathbf{G_{rr}} + G_{\mathrm{pr}} + \mathbf{G}_{\mathrm{qr}}$$

Direct links in G_{rr}

Indirect links in G_{qrnd}



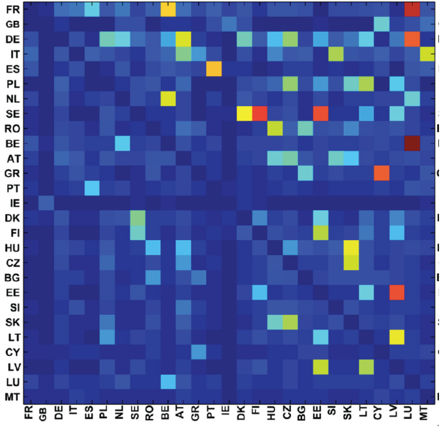


Innovative information in G_{qrnd} (non diagonal terms of G_{qr})

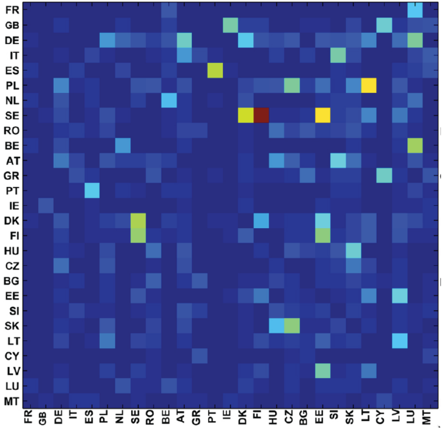
27 EU countries for EnWiki vs FrWiki

Cultural views

G_{qrnd} for EnWiki



G_{qrnd} for FrWiki

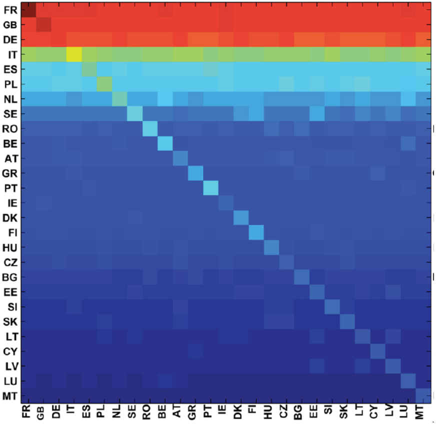


Networks of ‘friends’

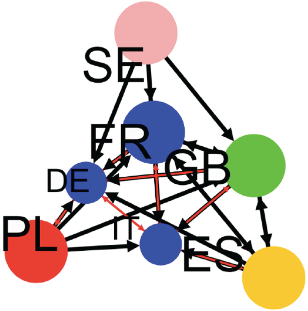
Top friends of country *j*

Ranking of countries by descending value of column *j*.

G_{R} for EnWiki



Top 4 friends network



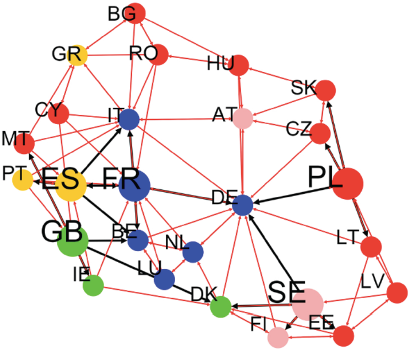
for 5 selected countries

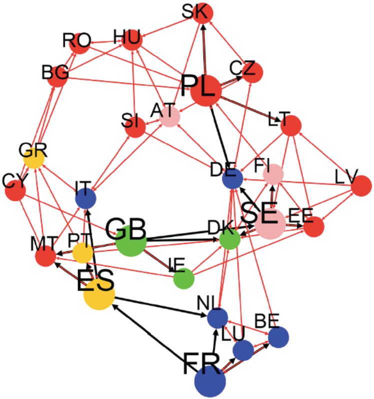
→ Dominated by PageRank

Networks of ‘friends’

Top ‘hidden’ friends from G_{qrnd}

Plotted automatically with a force-direct layout





Top 4 for EnWiki

Top 4 for FrWiki

Networks of ‘friends’

Cross-edition friendship

Friendship relations visible in all 5 editions (EnWiki, FrWiki, RuWiki, DeWiki, ArWiki)

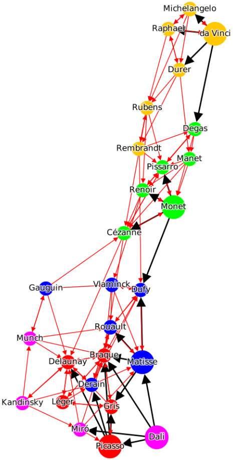
Selected country	G_{qr} Wiki friends present in		
	all 5 editions	4 out of 5 editions	3 out of 5 editions
FR	BE -ES	IT	
GB	IE		DK - FR
ES	IT - PT	FR	BE
SE	DK - FI		EE
PL	CZ		DE - HU - LT - SK

Networks of 30 Painters

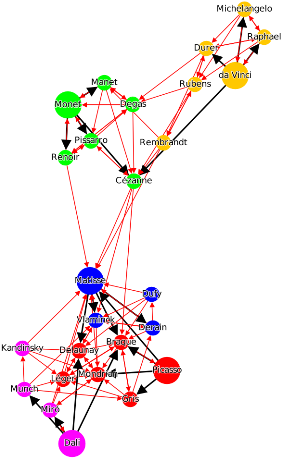
Name	Category	Colour	FrWiki	EnWiki	DeWiki
Picasso	**Cubism**	Red	1	2	2
Braque		Red	17	20	20
Léger		Red	19	24	24
Mondrian		Red	25	22	22
Gris		Red	29	28	25
Delaunay		Red	28	27	26
Matisse	**Fauvism**	Blue	6	11	12
Gauguin		Blue	13	15	18
Derain		Blue	22	25	27
Dufy		Blue	27	26	29
Rouault		Blue	30	30	28
Vlaminck		Blue	24	29	30
Monet	**Impressionists**	Green	4	9	11
Cézanne		Green	8	12	9
Manet		Green	12	13	16
Renoir		Green	15	14	17
Degas		Green	18	16	21
Pissarro		Green	23	19	23
da Vinci	**Great masters**	Orange	2	1	1
Michelangelo		Orange	3	3	4
Raphael		Orange	5	4	5
Rembrandt		Orange	9	5	6
Rubens		Orange	10	7	7
Durer		Orange	14	8	3
Dali	**Modern 20-21**	Pink	7	10	13
Warhol		Pink	11	6	8
Kandinsky		Pink	20	17	10
Chagall		Pink	21	18	15
Miró		Pink	16	21	19
Munch		Pink	26	23	14

Networks of Painters

For EnWiki



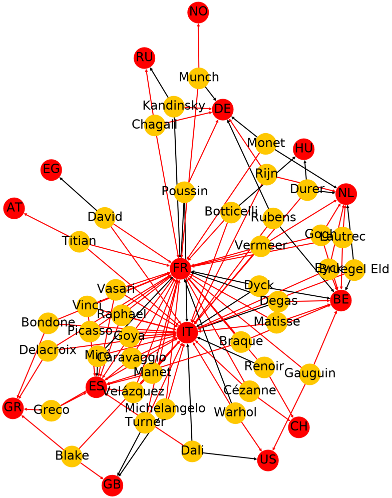
For FrWiki



Orange → Great masters ; Green → Impressionism ; Blue → Fauvism ;
Red → Cubism ; Pink → Modern (20-21).

Interaction between painters and countries

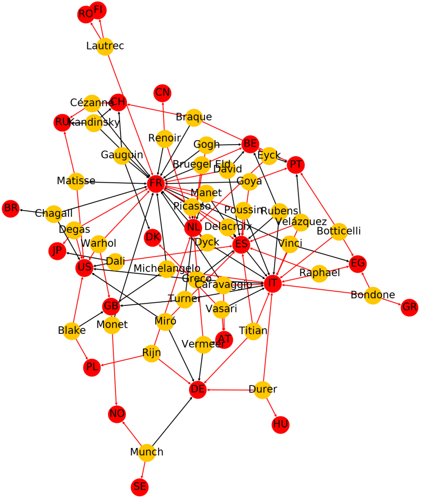
Top 3 country friends for top 40 painters network from $G_{rr} + G_{\mathrm{qrnd}}$ for EnWiki.



Black arrows : $G_{rr}(i,j) > G_{\mathrm{qrnd}}(i,j)$; Red arrows $G_{rr}(i,j) \le G_{\mathrm{qrnd}}(i,j)$.

Interaction between painters and countries

Top 3 country friends for top 40 painters network from $G_{rr} + G_{\mathrm{qrnd}}$ for **FrWiki**.



Black arrows : $G_{rr}(i,j) > G_{\mathrm{qrnd}}(i,j)$; Red arrows $G_{rr}(i,j) \leq G_{\mathrm{qrnd}}(i,j)$.

Sensitivity analysis

How does a relative link variation impact the reduced network structure?

Sensitivity analysis

For the relationship from nation $j \to i$ in G_{R}

- Modify element $\tilde{G}_{\mathrm{R}}(i,j) = (1+\delta) G_{\mathrm{R}}(i,j)$
- Normalize column j of $\tilde{G}_{\mathrm{R}}$.
- Calculate modified PageRank $\tilde{P}$ with $\tilde{G}_{\mathrm{R}}$.
  We observe a change of importance of nodes in the network.
- Calculate the logarithmic derivative of the PageRank probability of a given node k:

$$D_{(j\to i)}(k) = (\mathrm{d}P_k / \mathrm{d}\delta_{ij}) / P_k = (\tilde{P}_k - P_k)/(\delta_{ij} P_k)$$

This measures the **sensitivity of nation k to the link $j \to i$**.

27 EU countries average sensitivity

Average sensitivity across editions

Following sensitivity results $\bar{D}$ are averaged over 3 editions: EnWiki, FrWiki and DeWiki

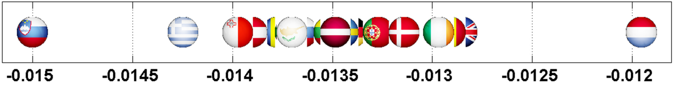


Figure: **Axial representation of $\bar{D}$ for a link modification from $\{$IT$\}$ to $\{$FR$\}$.** Here $\bar{D}(IT) = -0.0159$ and $\bar{D}(FR) = 0.0701$ are not represented.

Slovenia is mostly hit by an increase of Italy → France link

40 worldwide countries average sensitivity

Impact of China → US

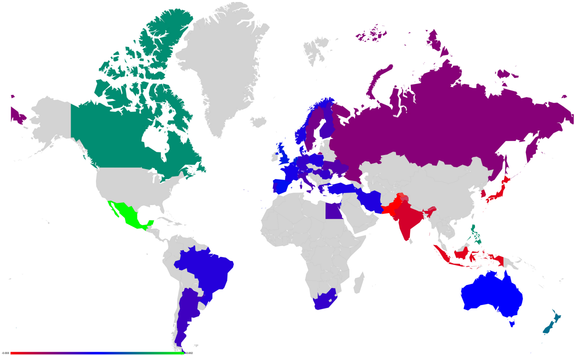


Figure: **Map representation of $\bar{D}$ for link modification from CN to US.** Non represented values: $\bar{D}(CN) = -0.0056$, $\bar{D}(US) = 0.0210$ and $\bar{D}(TW) = -0.0087$. Lower values in red, median blue and larger in green

40 worldwide countries average sensitivity

Clusters of countries

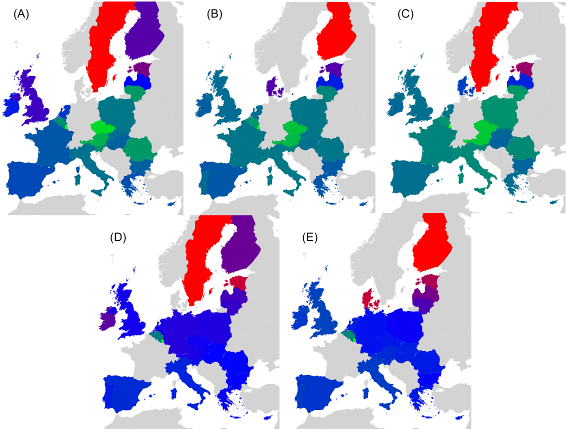
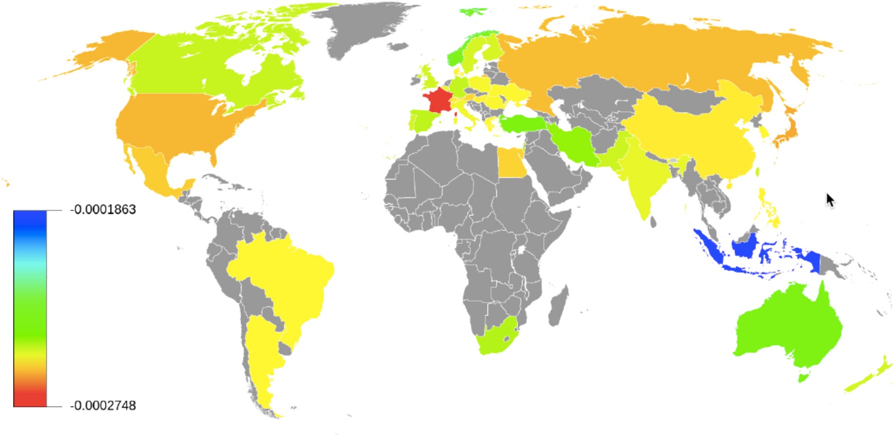


Figure: **Map representation of $\bar{D}$ for link modifications from Nordic countries to $\{$FR or DE$\}$.** (A): DK to DE. (B): SE to DE. (C): FI to DE. (D): DK to FR. (E): SE to FR.

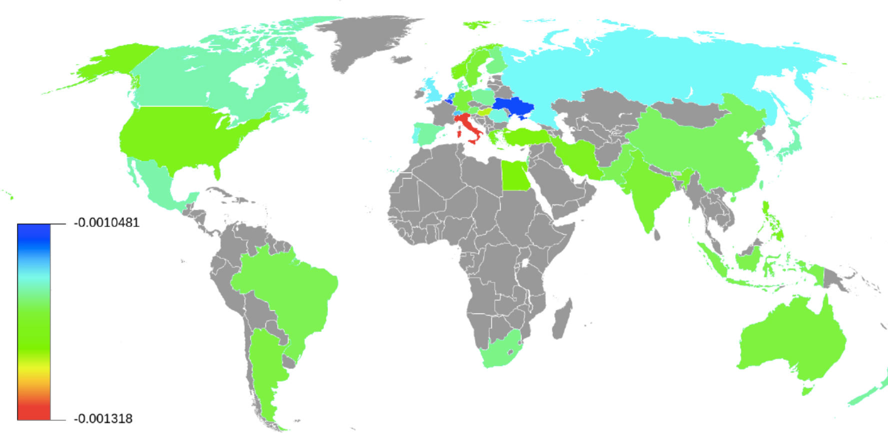
Average sensitivity of countries to painters

Sensitivity from Van Gogh → Netherlands over 40 countries.



Average sensitivity of countries to painters

Sensitivity from Da Vinci → France over 40 countries.



2-way sensitivity

Sensitivity to bi-directional changes

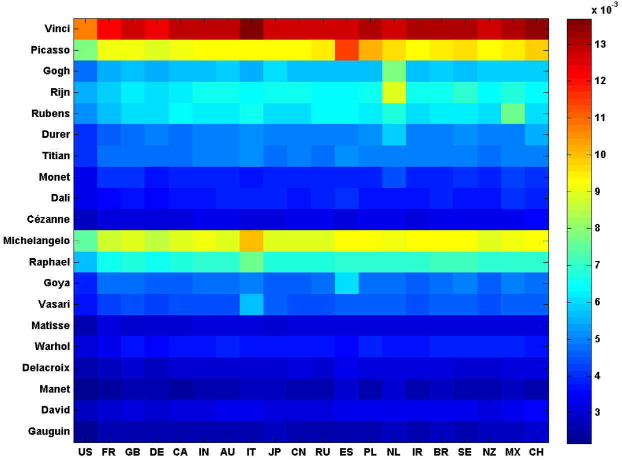
Measures the sensitivity of a nation *a* to the changes in both directions of link $i \to j$:

$$D_{(i \leftrightarrow j)}(a) = D_{(i \to j)}(a) + D_{(j \to i)}(a) \tag{2}$$

2-way sensitivity for painters and countries

Diagonal sensitivity of top 20 countries of 20 top painters

This is the 2-way sensitivity calculated for the top 20 countries when interaction is studied with top 20 painters.



Relationship imbalance analysis

Relationship imbalance between two nations

The 2-way sensitivity can help us know which country has the most influence on the other one.
For countries a and b we define:

$$F(a, b) = D_{(a \leftrightarrow b)}(a) - D_{(a \leftrightarrow b)}(b) \tag{3}$$

- $F(a, b) > 0$, b is the strongest nation
- $F(a, b) < 0$, a is the strongest nation

Relationship imbalance analysis

For 27 EU network

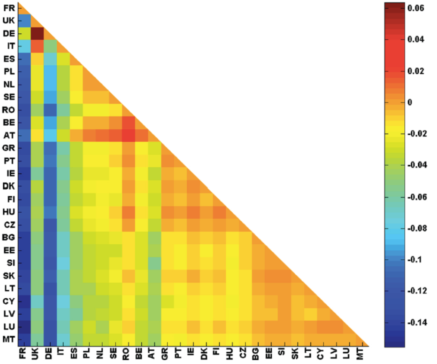
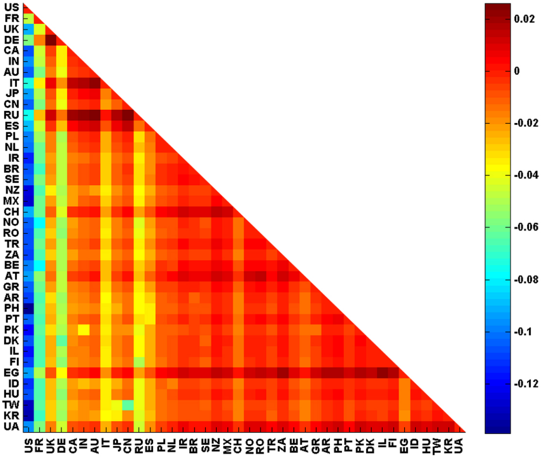


Figure: **Relationship imbalance analysis: F-representation for 27 EU network.** X-axis and Y-axis represent *a* and *b* respectively. If $F(a, b)$ is negative, nation *a* has more influence on nation *b* than *b* on *a*.

Relationship imbalance analysis

For 40 worldwide network



Conclusions

Google matrix analysis of Wikipedia

Offers a nice framework to automatically learn embedded information:

- Importance of nodes with PageRank and derivated metrics
- Exhibit interactions within a sub-network (thematic view) with Reduced Google matrix
- Understand the influence of links and nodes on the network with the sensitivity analysis.

Conclusions

Google matrix analysis of Wikipedia

Very nice properties to become a major tool for Artificial Intelligence and automatic information extraction.

Therefore, we have to:

- Automatically extract the articles that can constitute a good sub-network for a given study.
- Capture easily the evolution of the reduced network for a change of topology of the complete network.

Related publications

1. S. E. Zant, K. Jaffrès-Runser, K. M. Frahm, and D. Shepelyansky, ”Interactions and influence of world painters from reduced Google matrix of Wikipedia networks” in IEEE Access, vol. 6, pp. 47735-47750, August 2018.

2. S. E. Zant, K. Jaffrès-Runser, and D. Shepelyansky, ”Capturing the influence of geopolitical ties from Wikipedia with reduced Google matrix”, PLOS ONE 13(8), pp. 1-31, August 2018

3. S. E. Zant, K. M. Frahm, K. Jaffrès-Runser, and D. Shepelyansky, ”Analysis of world terror networks from the reduced Google matrix of Wikipedia” Springer, EPJB, vol. 91, no.1, pp. 7, January 2018.

4. K. M. Frahm, S. E. Zant, K. Jaffrès-Runser, and D. L. Shepelyansky, ”Multi-cultural Wikipedia mining of geopolitics interactions leveraging reduced Google matrix analysis” Elsevier, PLA, vol. 381, no. 33, pp. 2677 - 2685, September 2017.