palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

Received 7 Apr 2015 | Accepted 16 Nov 2015 | Published 15 Dec 2015

DOI: 10.1057/palcomms.2015.41

OPEN

Mapping bilateral information interests using the activity of Wikipedia editors

Fariba Karimi[1], Ludvig Bohlin[2], Anna Samoilenko[1], Martin Rosvall[2] and Andrea Lancichinetti[2]

ABSTRACT We live in a global village where electronic communication has eliminated the geographical barriers of information exchange. The road is now open to worldwide convergence of information interests, shared values and understanding. Nevertheless, interests still vary between countries around the world. This raises important questions about what today's world map of information interests actually looks like and what factors cause the barriers of information exchange between countries. To quantitatively construct a world map of information interests, we devise a scalable statistical model that identifies countries with similar information interests and measures the countries' bilateral similarities. From the similarities we connect countries in a global network and find that countries can be mapped into 18 clusters with similar information interests. Through regression we find that language and religion best explain the strength of the bilateral ties and formation of clusters. Our findings provide a quantitative basis for further studies to better understand the complex interplay between shared interests and conflict on a global scale. The methodology can also be extended to track changes over time and capture important trends in global information exchange.

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany [2] Department of Physics, Umea University, Umea, Sweden
Correspondence: (email: andrea.lancichinetti@umu.se)

Introduction

"We live in a global world" has become a cliché (Kose and Ozturk, 2014). Historically, the exchange of goods, money and information was naturally limited to nearby locations, since globalisation was effectively blocked by spatial, territorial and cultural barriers (Cairncross, 2001). Today, new technology is overcoming these barriers and exchange can take place in an increasingly international arena (Friedman, 2000). Nevertheless, geographical proximity still seems to be important for the trade of goods (Overman *et al.*, 2003; Serrano *et al.*, 2007; Fagiolo *et al.*, 2010; Kaluza *et al.*, 2010) as well as for mobile phone communication (Lambiotte *et al.*, 2008) and scientific collaboration (Pan *et al.*, 2012). However, since the Internet allows information to travel more easily and rapidly than goods, it remains unclear what are the effective barriers of global information exchange. As information exchange requires shared interests, we therefore need to better understand global connections in interest, and the factors that form these connections.

Although globalisation of information has been discussed extensively in the research literature (Friedman, 2000; Fischer, 2003; Nye, 2004), currently there is no method to quantitatively map bilateral information interests from large-scale data. Without such a method, it becomes difficult to justify qualitative statements about, for example, the complex interplay between shared values and conflict on a global scale. We use data mining and statistical analysis to device a measure of bilateral information interests, and use this measure to construct a world map of information interests.

To study interests on a global scale, we use the free online encyclopedia Wikipedia, which has evolved into one of the largest collaborative repositories of information in the history of mankind (Mesgari *et al.*, 2014). The free online encyclopedia consists of almost 300 language editions, with English being the largest one (http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia. This multi-lingual encyclopedia captures a wide spectrum of information in millions of articles. These articles undergo a peer-reviewed editing process without a central editing authority. Instead, articles are written, reviewed and edited by the public. Each article edit is recorded, along with a time-stamp, and, if the editor is unregistered, the computer's IP address. The IP address makes it possible to connect each edit to a specific location. Therefore we can use Wikipedia editors as sensors for mapping information interest to specific countries.

In this paper, we use co-editing of the same Wikipedia article as a proxy for shared information interests. To find global connections, we look at how often editors from different countries co-edit the same articles. To infer connections of shared interest between countries, we develop a statistical model and represent significant correlations between countries as links in a global network. Structural analysis of the network suggests that interests are polarised by factors related to geographical proximity, language, religion and historical background. We quantify the effects of these factors using regression analysis and find that information exchange indeed is constrained by the impact of social and economic factors connected to shared interests.

Methodology

Relating information interests to geographical location. As one of the largest and most linguistically diverse repositories of human knowledge, Wikipedia has become the world's main platform for archiving factual information (Mesgari *et al.*, 2014). One important feature of Wikipedia is that every edit made to an article is recorded. Thanks to this detailed data, Wikipedia provides a unique platform for studying different aspects of information processes, for example, semantic relatedness of topics (Auer and Lehmann, 2007; Radinsky *et al.*, 2011), collaboration (Kimmons, 2011; Keegan *et al.*, 2012; Török *et al.*, 2013), social roles of editors (Welser *et al.*, 2011) and the geographical locations of Wikipedia editors (Lieberman and Lin, 2009).

In this work, we used data from Wikipedia dumps (http://dumps.wikimedia.org/enwiki/) to select a random sample from the English Wikipedia edition, which is the largest and most widespread language edition. In total, the English edition has around 10 million articles, including redirects and duplicates. Since retrieving the editing histories of all articles is computationally demanding, we randomly sampled more than 6 million articles from this set. For each English article, we retrieved the complete editing history of the same article in all language editions that the English Wikipedia page links to. Finally we merged all language editions together to create a global editing history for each article. For each edit, the editing history includes the text of the edit, its time-stamp and, for unregistered editors, the IP address of the editor's computer. From the IP address associated with the edit, we retrieved the geolocation of the corresponding editor using an IP database (http://www.ip2location.com/). For the purpose of spatial analysis, we limited the analysis to edits from unregistered editors, because data on the location for most of the registered Wikipedia editors are unavailable. The resulting dataset contains more than 6 million (6,285,753) Wikipedia articles and about 140 million edits in total. We use these edits to create interest profiles for countries.

Inferring shared interests from edit co-occurrence. We identify the interest profile of a country by aggregating the edits of all Wikipedia editors whose IPs are recorded in the country. If an article is co-edited by editors located in different countries, we say that the countries share a common interest in the information of the article. In other words, we connect countries if their editors co-edit the same articles. Indirectly, we let individuals who edit Wikipedia represent the population of their country. While Wikipedia editors in a country certainly do not represent a statistically unbiased sample, there is a higher tendency that they edit contents that are related to the country in which they live (Hecht and Gergle, 2010b). Therefore, we approximate the interest profile of a country with collective editing behaviour of editors in that country.

Inferring the location of all editors on the country level is non-trivial. Although we have data on all edits, we do not know the location of registered editors because their IPs are not recorded. One proposed approach to tackle this problem makes use of circadian rhythms of editing activity to infer the location of the editors (Yasseri *et al.*, 2012). This method approximates the longitude of a location but provides little information about its latitude. Therefore, we must limit the analysis to the activity of unregistered editors with recorded IP addresses. This will arguably affect the results. Not only do registered editors contribute to 70% of all 140 million edits, they also have somewhat different behaviour. For example, many of the most active registered users take on administrative functions, develop career paths or specialise in covering selected topics (Arazy *et al.*, 2015). On the other hand, some unregistered editors are involved in vandalism, but often their activity nevertheless indicates their interest.[1] While we can only speculate about how including registered editors would affect the results, unregistered editors can nevertheless provide useful information about shared interests between countries.

From the co-editing data, we create a network that represents countries as nodes and shared interests as links. The naive approach is to use the raw counts of co-edits between countries

as weighted links. The problem with this approach is that it is biased towards the number of editors in each country. Some countries may be strongly connected, not because of evident shared interests but merely as a result of a large community of active Wikipedia editors. To address this problem, we propose a statistical validation method that filters out connections that could exist only due to size effects or noise. The filtering method assumes a multinomial distribution and determines the expected number of co-occurring edits from the empirical data. In other words, we infer significant links in a bipartite system in which countries are in one set and articles are in the other set. There are other existing methods to evaluate the significant correlation between entities in bipartite systems. For example, Zweig and Kaufmann (2011) proposed a systematic approach to one-mode projections of bipartite graphs for different motifs. In another work, Tumminello *et al.* (2011) used the hypergeometric distribution and measured the *P*-value for each subset of the bipartite network. Moreover, Lancichinetti *et al.* (2015) proposed a community detection method to classify topics to articles more efficiently, and Serrano *et al.* (2009) used a disparity filtering method to infer significant weights in networks. Finally, Ronen *et al.* (2014) adopted a statistical approach to determine significant links between languages in various written documents. However, the model that we use has the advantages that it makes it easy to account for size variation inside an article and to compute the *z*-scores for analysing the country-based editor activity.

Interest model. In this section we outline the formalisation of the model. We link countries based on their co-occurring edits over all Wikipedia articles. For a specific article a, we calculate the link weight between all pairs of countries that edited the article, as follows: if editors in country i have edited an article k_i^a times, and editors in country j have edited the same article k_j^a times, then the countries' empirical link weight, w_{ij}^a, is calculated as:

$$w_{ij}^a = k_i^a k_j^a \tag{1}$$

Since the total number of articles is over 6 million, most country pairs have co-edited at least one article. Therefore, the aggregation of all articles results in numerous links between countries, and the countries with relatively large editing activities become highly central. Accordingly, we cannot know if the link exists by chance, or because countries actually tend to edit the same articles more frequently than expected. To determine which links are statistically significant, we compare the empirically observed link weights with the weight given by a null model. In the null model, we assume that each edit comes from a country randomly picked proportionally to its total number of edits. More specifically, the random assignments are performed by drawing the countries from a multinomial distribution. That is, for each edit, country i is selected proportional to its cumulative editing activity, $p_i = \sum_a k_i^a / M$, where M is the total number of edits for all articles. Note that each edit is sampled independently from all other edits, and that the cumulative edit activity of a country in the null model on average will be the same as the observed one. This null model preserves the average level of activity of the countries, but randomises the temporal order and the articles that countries edit. Figure 1 shows an example of this reshuffling scheme with four articles.

From the null model, we can analytically compute the expected probability, μ_{ij}^a, that two countries i and j edit the same article a (detailed derivation in the Supplementary Information S1):

$$\mu_{ij}^a = n_a(n_a - 1)p_i p_j \tag{2}$$

where n_a is the total number of edits in article a.

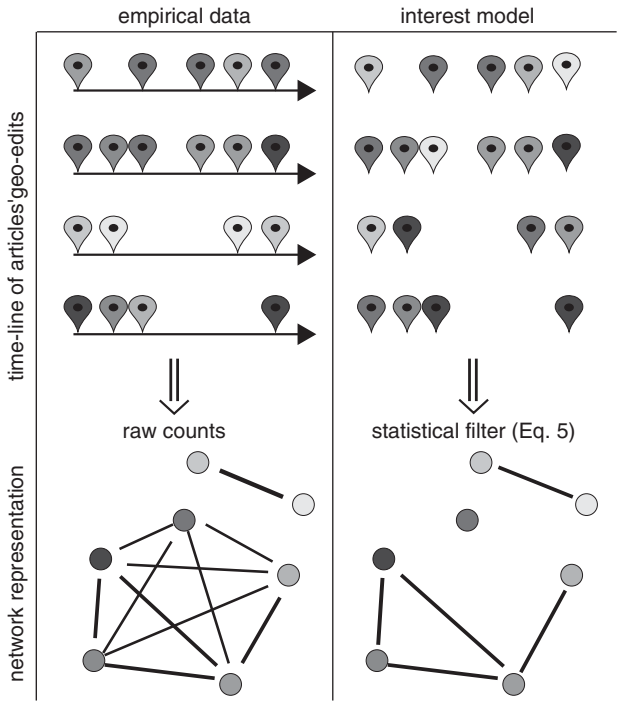

Figure 1 | Illustration of the interest model.
***Notes*: The left panel shows the time-ordered edit sequence of four Wikipedia articles, with edits coming from different countries represented as coloured pins. Note that pins represent country edits, and therefore they can be repeated. The resulting empirical network, calculated by multiplying raw co-edits counts, is at the bottom. In the right panel, we illustrate the null model with the same four articles, and the resulting network at the bottom. In the null model, the average editing activity of the countries is preserved, but the order of the edits is reshuffled within and across articles. As a result of the filtering, some links are removed in the interest model.**

To compare the empirical and expected link values, we compute standardised values, so called *z*-scores. For countries i and j and article a, the *z*-score z_{ij}^a is defined as

$$z_{ij}^a = \frac{w_{ij}^a - \mu_{ij}^a}{\sigma_{ij}^a} \tag{3}$$

where the standard deviation σ_{ij}^a, is computed in the Supplementary Information S1.

The *z*-scores are useful for comparisons of weights, since they account for the large variations that exist in the articles' edit histories. We then sum over all articles to find the cumulative *z*-score for countries i and j

$$z_{ij} = \sum_a \frac{w_{ij}^a - \mu_{ij}^a}{\sigma_{ij}^a} \tag{4}$$

Using the Bonferroni correction, we consider a link to be significant if the probability of observing the total *z*-score is less than $0.05/N$, where N is the number of countries. Since the total *z*-score is a sum over many independent variables, we can approximate the expected total *z*-score distribution with a normal distribution. The normal distribution has average value 0 and standard deviation $\sqrt{L}$, where L is the number of Wikipedia articles. Thus, the threshold for the significant link weight is $t = 3.52\sqrt{L}$, where 3.52 is derived from the condition that $P(z > 3.52) = 0.05/N$, where $N = 234$ countries and P is the standard Gaussian distribution (with zero average and unit variance). If the total

Figure 2 | World map of information interests.
***Notes*: Countries that belong to the same cluster have the same colour. Countries coloured with gray do not belong to any cluster. The map suggests that the division of countries can be explained by a combination of cultural and geopolitical features.**

z-score is larger than the threshold, we create a link between countries i and j with weight $\tilde{w}_{ij}$ according to

$$\tilde{w}_{ij} = \begin{cases} z_{ij} - t & \text{if } z_{ij} > t \\ 0 & \text{if } z_{ij} \leqslant t \end{cases} \quad (5)$$

In summary, the interest model maintains the average level of activity of the countries and randomises the articles that they edit. By comparing results from the interests model and empirical values, we can identify significant links between countries.

Clustering countries with similar interests. To investigate the effective barriers of global information exchange, we first identify large-scale structures among the thousands of links between countries. In this way, we can highlight the groups of countries that share interest in the same information. To reveal such groups among the pairwise connections, we use a network community detection method based on random walks as a proxy for interest flows. While the community-detection method we use is good at breaking chains of links, we may connect some countries primarily based on strong connections with common countries and not between themselves. Nevertheless, simplifying and highlighting important structures provide a valuable map to investigate the large-scale structure of global information exchange.

In our clustering approach, we first build a network of countries connected with the significant links we find in our filtering. To identify groups of countries, we envision an editor game in which editors from different countries are active in sequence. In this relay race, a country exchanges information to another country proportional to the weight of the link between the countries. Accordingly, the sequence of countries forms a random walk and certain sets of countries with strong internal connections will be visited for a relatively long time. This process is analog to the community-detection method known as the map equation (Rosvall and Bergstrom, 2008; Rosvall *et al.*, 2009). Here we use the map equation's associated search algorithm Infomap (Edler and Rosvall, 2015) to identify the groups of countries we are looking for and to reveal the large-scale structure.

Results and discussion

We will discuss the results at four levels of detail, from the big picture to the detailed dynamics, and highlight different potential mechanisms for barriers of information exchange. First, we will show a global map of countries with shared information interests, and continue with the interconnections between the clusters. Then we will consider each cluster separately and examine the interconnections between countries within the clusters. Finally, we will apply multiple regression analysis to examine explanatory variables to the highways and barriers of information exchange.

The world map of information interests. The world map of information interests suggests that cultural and geopolitical features can explain the division of countries. Between the 234 countries, we identified 2,847 significant links that together form a network of article co-edits. By clustering the network, we identified 18 clusters of strongly connected countries (see Supplementary Table for a detailed list of countries in each cluster). The resulting network is illustrated as a map in Fig. 2, where countries of the same cluster share the same colour. The map suggests that the division of countries can be explained by a combination of cultural and geopolitical features. For example, the United States and Canada share a long geographical border and extensive mutual trade, and are clustered together despite the fact that other English-speaking countries are not. Moreover, religion is a plausible driver for the formation of the cluster of

countries in the Middle East and North Africa, as well as the cluster of Russia and the Orthodox Eastern-European countries (Gupta *et al.*, 2002). Another factor in the formation of shared information interests is language. For example, countries in Central and South America are divided into two clusters with Portuguese and Spanish as common languages in each cluster, respectively. Colonial history can also shape similarity in interests, as in the cluster of Portugal, Angola and Brazil, as well as the cluster of former Soviet Union countries (Hensel, 2009). Overall, there is strong empirical evidence that geographical proximity, common religion, shared language and colonial history can explain the division of countries.

To examine the connections between clusters, we looked at the network structure at the cluster level. The network in Fig. 3 shows the connections between the clusters of countries illustrated in Fig. 2 with the same colour coding. Connections tend to be stronger between clusters of geographically proximate countries also at this level. Interestingly, the Middle East cluster in turquoise has the largest link strengths to other clusters, forming a hub that connects East and West, North and South. Interpreting the strong connections as potential highways for information exchange, the Middle East is not only a melting pot of ideas, but also plays an important role in the spread of information.

To get better insights into how the clusters are shaped, we zoomed into the country networks within clusters. In the upper left corner of Fig. 3, we show the strongest connections within the Central European cluster. It suggests that countries are linked based on the overlap of the official languages (Hale, 2014). For example, Belgium has three official languages, Dutch, French and German. Indeed, Belgium is connected closely with the Netherlands, France and Luxembourg. We observed the same pattern in other clusters, and the triad of Switzerland, Germany and Austria is another example of strongly linked countries with a shared language.

Just to illustrate what interests can form the bilateral connections, we looked at a number of concrete examples. First, we ranked the articles according to their significant *z*-scores for each pair of countries. Then we looked at the top-ranked articles and here report the results for two European country pairs: Germany—Austria in the European cluster and Sweden—Norway in the Scandinavian cluster. The articles with the most significant co-edits relate to local and regional interests, including sports, media, music and places (see Supplementary Table). For example, the top-ranked articles in the Germany—Austria list include an Austrian singer who is also popular in Germany, and an Austrian football player who is playing in the German league. The top-ranked articles in the Sweden–Norway list shows a similar pattern of locally related topics, for example, a host of a popular TV show simultaneously aired in Sweden and Norway, a Swedish football manager who has been successful both in Sweden and Norway, and a music genre that is nearly exclusive to Scandinavian countries. Altogether, the top articles suggest that an important factor for co-editing is related interests, which in turn may be an effect of shared language, religion, or colonial history, as well as geographical proximity or large volume of trade between countries.

Regression analysis of the highways and barriers of information exchange. To quantify the impact of social and economic factors behind the shared interests, we performed a Multiple

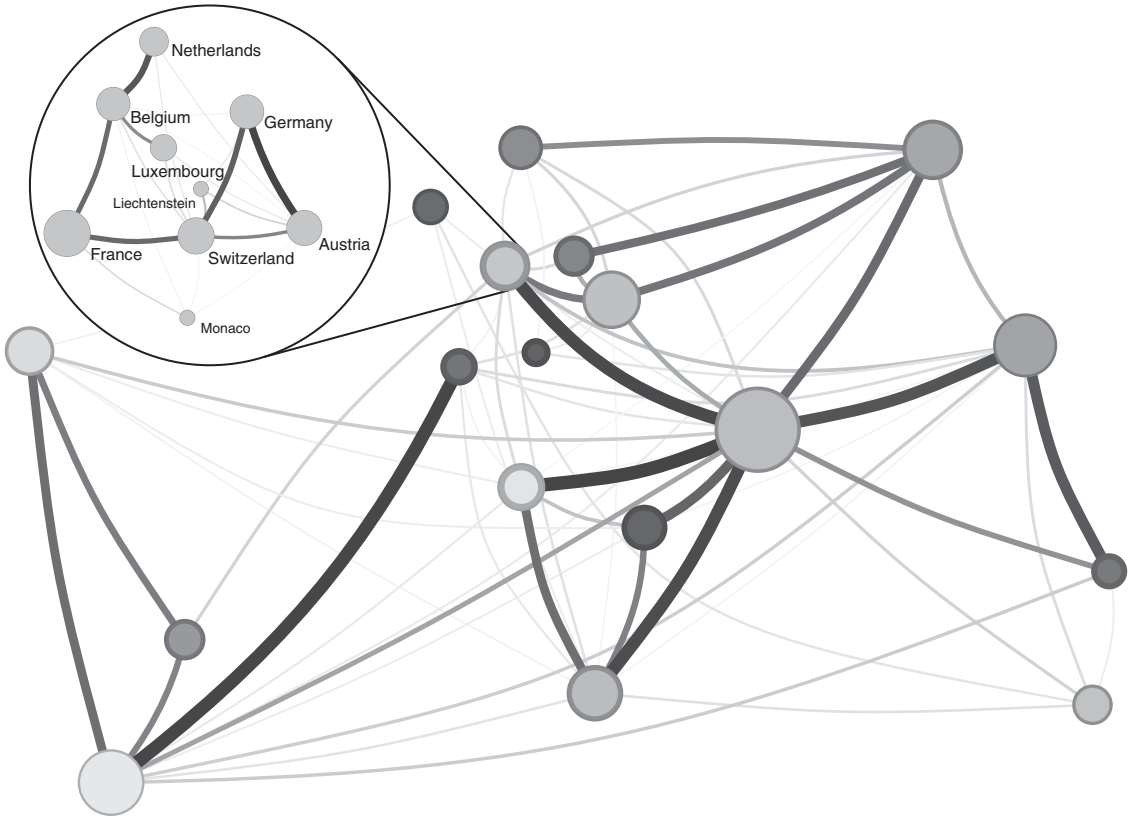

Figure 3 | World network of information interests.
***Notes*: The size of the nodes represents the total *z*-score of the clusters. The links represent connections between clusters obtained from the cluster analysis with Infomap, the thicker the line, the stronger the connection. Clusters are coloured in the same way as in Fig. 2. The upper left corner shows the most significant connections between countries in the Central European cluster.**

Table 1 | Results of the multiple regression analysis

	R_0	$R1$	$R2$	$R3$	$R4$
Intercept	0.41	0.3	2.33	2.33	2.28
Shared language	0.91* (69)	0.82* (64)	0.77* (60)	0.75* (58)	0.74* (57)
Shared religion		2.76* (46)	2.6* (44)	2.6* (43)	2.44* (40)
Log distance			−0.23* (−23)	−0.23* (−23)	−0.23* (−23)
Colonial tie				4.5* (22)	4.35* (21)
Log trade					0.03* (10)
Adjusted R^2	0.13	0.19	0.20	0.21	0.22
F-statistic	7,774	3,590	2,610	2,110	1,716
dF	30,874	30,873	30,872	30,871	30,870

Notes: Significant edit co-occurrences (z-scores) form the dependent variable matrix, which we regress on the independent matrices in different models. Values in parenthesis are t-statistics. The features are ordered by importance, from shared language to trade. Country pairs = 62,001. Values marked with an asterisk have a P-value less than 0.01

Regression Quadratic Assignment Procedure (MRQAP) analysis. This method is specifically suited when there are collinearity and autocorrelation in the data (Krackhardt, 1988; Dekker *et al.*, 2007). We performed the MRQAP using the netlm function in the sna R package (Butts, 2008). The dependent variables in the regression model were the significant z-scores that we obtained from the data. Our independent variables were geographical proximity, trade (Subramanian and Wei, 2007), colonial ties,[2] shared language[3] and shared religion,[4] as suggested by the analysis of the map of information interests (see the Supplementary Information S2 for a more detailed description of the data).

All independent variables show significant correlation with the data (see Table 1). To observe the variation between different independent matrices, we examined them in different models. In model R_0, we examined the influence of shared language, which explains 13% of our observation. In model R_1, we added shared religion, which increases the power of the model to 19%. In model R_2, we included the geographical proximity. It slightly increases the R^2 and has negative relation to the observed z-scores, since short distance corresponds to high proximity. In models R_3 and R_4, respectively, we added colonial ties and trade. Including all these explanatory variables into the regression model enabled us to increase the explanatory power of the model to 22%. The correlation of each variable with the observed z-scores can be inferred from the t-statistic. Shared language shows the strongest association, followed by shared religion, geographical proximity, colonial ties, and volume of trade (see Table 1).

The influence of language on shared interests is not surprising. It is well known that interests are formed by cultural expression and public opinion, and language is an important platform for these expressions (Usunier and Lee, 2005). That the relation between language and interests is important has also been demonstrated by the surprisingly small overlap between languages in Wikipedia (Hecht and Gergle, 2010a; Callahan and Herring, 2011) and the variation in the editing of controversial topics (Yasseri *et al.*, 2014).

Moreover, the influence of religion is in line with the Huntington's thesis that the source of division between people in the post-Cold War period is primarily rooted in cultural differences and religion (Huntington, 1997). Similar results were found in other studies that analysed Twitter and email communication worldwide (State *et al.*, 2015).

Overall, the analysis reveals that information exchange is constrained by the impact of social and economic factors connected to shared interests. In other words, globalisation of the technology does not bring globalisation of the information and interests. Language, religion, geographical proximity, historic background and trade are potential driving factors to polarise the information interests. These results coincide with earlier works that highlight the impact of the colonisation, immigration, economics, and politics on the cultural similarities and diversities (Tägil, 1995; Feldman-Bianco, 2001; Risse, 2001; Bleich, 2005; Castells, 2011; Gelfand *et al.*, 2011; Hennemann *et al.*, 2012).

Conclusions

By simplifying and highlighting the important structure in the myriad edits of Wikipedia, we provide a world map of shared information interests. We find that information interests despite globalisation are diverse, and that the highways and barriers of information exchange are formed by social and economic factors connected to shared interests. In descending order, we find that language, religion, geographical proximity, historic background and trade explain the diversity of interests. While technological advances in principle have made it possible to communicate with anyone in the world, these social and economic factors limit us from doing so and information interests remain diverse. Questions remain how different social and economic factors affect different regions, how they relate to conflicts on a global scale, and how the impact of these factors changes over time. It would therefore be interesting to extend the methodology to track changes over time.

Notes

1 See Wikipedia's policy and fight against vandalism here: https://en.wikipedia.org/wiki/Vandalism_on_Wikipedia.
2 We used ICOW Colonial History Data Set, version 1.0. available on http://www.paulhensel.org/icowcol.html.
3 We used Ethnologue available on http://www.ethnologue.com/.
4 We used the World Religion Database available on http://www.worldreligiondatabase.org/.

References

Arazy O, Ortega F, Nov O, Yeo L and Balila A (2015) Functional roles and career paths in Wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, pp 1092–1105.

Auer S and Lehmann J (2007) What have Innsbruck and Leipzig in common? extracting semantics from Wiki content. In *The Semantic Web: Research and Applications*. Springer, pp 503–517.

Bleich E (2005) The legacies of history? Colonization and immigrant integration in Britain and France. *Theory and Society*; **34** (2): 171–195.

Butts C T (2008) Social network analysis with sna. *Journal of Statistical Software*; **24** (6): 1–51.

Cairncross F (2001) *The Death Of Distance: How The Communications Revolution Is Changing Our Lives*. Harvard Business Press: Boston, MA.

Callahan E S and Herring S C (2011) Cultural bias in Wikipedia content on famous persons. *Journal of the Association for Information Science and Technology*; **62** (10): 1899–1915.

Castells M (2011) *The Power Of Identity: The Information Age: Economy, Society, And Culture*. Volume 2. Chichester, UK: John Wiley & Sons.

Dekker D, Krackhardt D and Snijders T A (2007) Sensitivity of mrqap tests to collinearity and autocorrelation conditions. *Psychometrika*; **72** (4): 563–581.
Edler D and Rosvall M (2015) The Infomap software package, http://www.mapequation.org, accessed 1 May 2015.
Fagiolo G, Reyes J and Schiavo S (2010) The evolution of the world trade web: A weighted-network analysis. *Journal of Evolutionary Economics*; **20** (4): 479–514.
Feldman-Bianco B (2001) Brazilians in Portugal, Portuguese in Brazil: Constructions of sameness and difference 1. *Identities. Global Studies in Culture and Power*; **8** (4): 607–650.
Fischer S (2003) Globalization and its challenges. *American Economic Review*; **93** (2): 1–30.
Friedman T L (2000) *The Lexus And The Olive Tree: Understanding Globalization*. Palgrave Macmillan: London, UK.
Gelfand M J *et al* (2011) Differences between tight and loose cultures: A 33-nation study. *Science*; **332** (6033): 1100–1104.
Gupta V, Hanges P J and Dorfman P (2002) Cultural clusters: Methodology and findings. *Journal of world business*; **37** (1): 11–15.
Hale S A (2014) Multilinguals and Wikipedia editing. In *Proceedings of the Conference on Web Science*. ACM, pp 99–108.
Hecht B and Gergle D (2010a) The tower of babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp 291–300.
Hecht B J and Gergle D (2010b) On the "localness" of user-generated content. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pp 229–232. New York: ACM. ISBN 978-1-60558-7950. doi: 10.1145/1718918.1718962.
Hennemann S, Rybski D and Liefner I (2012) The myth of global science collaboration–collaboration patterns in epistemic communities. *Journal of Informetrics*; **6** (2): 217–225.
Hensel P R (2009) ICOW colonial history data set, version 0.4. *University of North Texas*, http://www.paulhensel.org/icowcol.html.
Huntington S P (1997) *The Clash of Civilizations and the Remaking of World Order*. Penguin Books: New Delhi, India.
Kaluza P, Kölzsch A, Gastner M T and Blasius B (2010) The complex network of global cargo ship movements. *Journal of the Royal Society Interface*; **7** (48): 1093–1103.
Keegan B, Gergle D and Contractor N (2012) Do editors or articles drive collaboration? Multilevel statistical network analysis of Wikipedia coauthorship. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM, pp 427–436.
Kimmons R M (2011) Understanding collaboration in Wikipedia. *First Monday*; **16** (12).
Kose M A and Ozturk E O (2014) A world of change. *Finance & Development*; **51**(3): 6–11.
Krackhardt D (1988) Predicting with networks: Nonparametric multiple regression analysis of dyadic data. *Social Networks*; **10** (4): 359–381.
Lambiotte R *et al* (2008) Geographical dispersal of mobile communication networks. *Physica A*; **387** (21): 5317–5325.
Lancichinetti A, Sirer M I, Wang J X, Acuna D, Körding K and Amaral L A N (2015) High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*; **5** (1): 011007.
Lieberman M D and Lin J (2009) You are where you edit: Locating Wikipedia contributors through edit histories. In: *Proceedings of the Third International ICWSM Conference*, San Jose, CA, pp 106–113.
Mesgari M, Okoli C, Mehdi M, Nielsen F Å and Lanamäki A (2014) "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*; **66** (2): 219–245.
Nye J S Jr (2004) *Power in the Global Information age: From Realism to Globalization*. Routledge: London, UK.
Overman H G, Redding S and Venables A (2003) *The Economic Geography Of Trade, Production and Income: A Survey Of Empirics*. Blackwell Publishing: Malden, MA.
Pan R K, Kaski K and Fortunato S (2012) World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports*; **2**: 902.
Radinsky K, Agichtein E, Gabrilovich E and Markovitch S (2011) A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the International Conference on World Wide Web*. ACM, pp 337–346.
Risse T (2001) A European identity? Europeanization and the evolution of nation-state identities In: Cowles M G, Caporaso J A and Risse-Kappen T (eds) *Transforming Europe: Europeanization and Domestic Change*. Cornell University Press: Ithaca, NY.
Ronen S, Gonçalves B, Hu K Z, Vespignani A, Pinker S and Hidalgo C A (2014) Links that speak: The global language network and its association with global fame. In *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.1410931111.
Rosvall M, Axelsson D and Bergstrom C T (2009) The map equation. *EPJ ST*; **178** (1): 13–23.
Rosvall M and Bergstrom C T (2008) Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*; **105** (4): 1118–1123.
Serrano M Á, Boguñá M and Vespignani A (2007) Patterns of dominant flows in the world trade web. *Journal of Economic Interaction and Coordination*; **2** (2): 111–124.
Serrano M Á, Boguñá M and Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*; **106** (16): 6483–6488.
State B, Park P, Weber I and Macy M (2015) The mesh of civilizations in the global network of digital communication. *PLoS ONE*; **10** (5): e0122543.
Subramanian A and Wei S J (2007) The WTO promotes trade, strongly but unevenly. *Journal of international Economics*; **72** (1): 151–175.
Tägil S (1995) *Ethnicity and Nation Building in the Nordic World*. SIU Press: London, UK.
Török J, Iñiguez G, Yasseri T, San Miguel M, Kaski K and Kertész J (2013) Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment. *Physical Review Letters*; **110** (8): 088701.
Tumminello M, Micciché S, Lillo F, Piilo J and Mantegna R N (2011) Statistically validated networks in bipartite complex systems. *PLoS ONE*; **6** (3): e17994.
Usunier J C and Lee J (2005) *Marketing across Cultures*. Pearson Education: Harlow, UK.
Welser H T *et al* (2011) Finding social roles in Wikipedia. In *Proceedings of the iConference*. ACM, pp 122–129.
Yasseri T, Spoerri A, Graham M, Kertész J (2014) The most controversial topics in Wikipedia: A multilingual and geographical analysis In: Fichman P and Hara N (eds) *Global Wikipedia: International and cross-cultural issues in online collaboration*. Rowman & Littlefield: Lanham, MD.
Yasseri T, Sumi R and Kertész J (2012) Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PloS ONE*; **7** (1): e30091.
Zweig K A and Kaufmann M (2011) A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining*; **1** (3): 187–218.Supplementary figure Supplementary Information

Data Availability

The datasets generated during and/or analysed during the current study are available in the Google Sites repository: https://sites.google.com/site/mappingbilateralwiki/.

Acknowledgements

The authors thank Alcides V. Esquivel, Daniel Edler, Claudia Wagner and Micheal Macy for valuable discussions. Authors also thank the Wikimedia Foundation for providing free access to the data. F.K., L.B., A.L. and M.R. were supported by the Swedish Research Council grant 2012-3729.

Additional Information

Supplementary Information: accompanies this paper at http://www.palgrave-journals.com/palcomms

Competing interests: The author declare no competing financial interests.

Reprints and permission information is available at http://www.palgrave-journals.com/pal/authors/rights_and_permissions.html

How to cite this article: Karimi, F. *et al* (2015) Mapping bilateral information interests using the activity of Wikipedia editors. *Palgrave Communications*. 1:15041 doi: 10.1057/palcomms.2015.41.