REGULAR ARTICLE   **Open Access**

CrossMark

Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity

Anna Samoilenko[1,2]*, Fariba Karimi[1], Daniel Edler[3], Jérôme Kunegis[2] and Markus Strohmaier[1,2]

*Correspondence:
anna.samoilenko@gesis.org
[1]GESIS - Leibniz-Institute for the
Social Sciences, 6-8 Unter
Sachsenhausen, Cologne, 50667,
Germany
[2]University of Koblenz-Landau,
Koblenz, Germany
Full list of author information is
available at the end of the article

Abstract

In this paper, we study the network of global interconnections between language communities, based on shared co-editing interests of Wikipedia editors, and show that although English is discussed as a potential *lingua franca* of the digital space, its domination disappears in the network of co-editing similarities, and instead local connections come to the forefront. Out of the hypotheses we explored, bilingualism, linguistic similarity of languages, and shared religion provide the best explanations for the similarity of interests between cultural communities. Population attraction and geographical proximity are also significant, but much weaker factors bringing communities together. In addition, we present an approach that allows for extracting significant cultural borders from editing activity of Wikipedia users, and comparing a set of hypotheses about the social mechanisms generating these borders. Our study sheds light on how culture is reflected in the collective process of archiving knowledge on Wikipedia, and demonstrates that cross-lingual interconnections on Wikipedia are not dominated by one powerful language. Our findings also raise some important policy questions for the Wikimedia Foundation.

Keywords: Wikipedia; multilingual; cultural similarity; network; digital language divide; socio-linguistics; digital humanities; hypothesis testing

1 Introduction

Measuring the extent to which cultural communities overlap via the knowledge they preserve can paint a picture of how culturally proximate or diverse they are. Wikipedia, the largest crowd-sourced encyclopedia today, is a platform that documents knowledge from different cultural communities via different language editions. The collective traces left by editors of Wikipedia can be utilized to identify cultural communities that are most similar with regard to the knowledge they document. Certainly, co-editing similarities among language communities of Wikipedia editors are just a particular dimension of culture and are not representative of cultural similarities among the communities in general. Yet, Wikipedia plays a critical role in today's information gathering and diffusion processes and Wikipedians constitute an important cultural subset of educated and technology-savvy elites who often drive the cultural, political, and economic processes [1]. In this paper, we tap into the traces left by editors of Wikipedia to gain new insights into how language communities on Wikipedia relate to each other via common co-editing interests.

Springer

Problem. We are thus interested in seeking answers to the following overarching research question: What are common editing interests between language communities on Wikipedia, and how can they be explained? In addition, we also aim to establish a computational method which would allow measuring culture-related similarities based on the topics the editors document in Wikipedia.

We assume that collective interest of a language-speaking community is reflected through the aggregation of articles documented in the corresponding language edition of Wikipedia. These articles are an approximation of the topics which are culturally relevant to that language community, though by no means are representative of the entire underlying cultural community. We define *cultural similarity* as a significant interest of communities in editing articles about the same topics; in other words, language communities are similar when they significantly agree regarding the topics they choose to edit.

Methods. Our approach consists of several steps. We first use statistical filtering to identify language pairs which show consistent interest in articles on the same topics. Based on this dyadic information, we create a network of interest similarity where nodes are languages and links are weighted as the strength of shared interest between them. We cluster the network and inspect it visually to inform the generation of hypotheses about the mechanisms that contribute to cultural similarity. Finally, we express these hypotheses as transition probability matrices, and test their plausibility using two statistical inference techniques - HypTrails [2] and MRQAP [3] (Multiple Regression Quadratic Assignment Procedure). Using both Bayesian and frequentist approaches, we obtain similar results, which suggests that our findings are robust against the chosen statistical measure.

Contribution and findings. Our main contribution is empirical. We expand the literature on culture-related research by (a) presenting a large-scale network of interest similarities between 110 language communities, (b) showing that the set of languages covering a concept on Wikipedia is not a random choice, and (c) by statistically demonstrating that similarity in concept sets between Wikipedia editions is influenced by multiple factors, including bilinguality, proximity of these languages, shared religion, and population attraction. We also combine multiple techniques from network theory, Bayesian and frequentist statistics in a novel way, and present a generalisable approach to quantify and explain culture-related similarity based on editing activity of Wikipedia editors.

We find that the topics that each language edition documents are not selected randomly, however small the underlying community of editors. We test several hypotheses about the underlying processes that might explain the observed nonrandomness, and find that bilingualism, linguistic similarity of languages, and shared religion provide the best explanations for the similarity of interests between cultural communities. Population attraction and geographical proximity are also significant, but much weaker factors bringing communities together.

The remainder of the paper is structured as follows. In Section 2 we will give a brief overview of work on how cultural differences find reflection in multilingual online platforms, as well as on how Wikipedia has been used to compare cultural and linguistic points of view, and cultural biases involved in knowledge production. In Section 3 we will describe in detail the process of data sampling and collection. Sections 4 and 5 will focus on identifying and explaining co-editing interests, give a technical overview of the quan-

titative methods, and report the results. We will offer our reflection upon the findings in Sections 6 and 7.

2 Related literature

Definition of culture and its borders is a long-debated and still unresolved issue in Anthropology and Social Sciences; a 1951 review of the works on the issue already contained close to 300 definitions of culture [4]. Cultural communities have fuzzy boundaries: several distinct cultures might co-exist in one state, or alternatively, reach beyond and across continents. This is especially true for multilingual countries or those with colonial past. While there are many non-verbal expressions of material culture, language is an important bearer of culture - its meanings have to be learnt socially and represent the way of life as seen by a particular community [5–8]. Language-speaking communities form distinct and unique cultures around themselves [9, 10], and overlap of interests between these communities might signify cultural proximity between them. Language is central to culture for several reasons: it reflects the collective agreement of a language community to view the world in a certain way, and helps a community to perpetuate its culture, develop its identity, and archive accumulated knowledge [11]. It is the latter feature of collective knowledge selection and archiving that this paper focuses on.

Wikipedia as a lens for studying cultural repertoires of language communities. The online encyclopedia Wikipedia is a prominent example of collective knowledge accumulation, and it is becoming one of the most interesting and convenient sources for academics to study cultural and historical processes [12]. Wikipedia is one of the most linguistically diverse projects online, with a constant base of editors contributing in almost 300 languages [13], ranging from almost 5 million in the largest edition (English) to just 89 in Cree, the smallest one [13]. This makes it accessible to more than 5 billion people, or 75% of the world's population [14]. There is no central authority that dictates which topics must be covered, and every editor is free to select their own, as long as they are consistent with the notability guidelines [15]. All language editions have their own notability guidelines and are edited independently from each other, although an editor can also co-edit several editions in parallel. Large language editions like English are not supersets of smaller ones, and each edition contains unique concepts which are not covered by others. For example, concept overlap between the two largest editions, English and German, is only 51% [16]. Opposite to the common misconception, even when articles on the same concept exist in different language editions, they are not translated replicas of each other, but instead reveal consistent cultural biases [17, 18] and introduce various linguistic viewpoints [19–21].

These differences in number, selection, and content of articles across languages are not accidental, but relate to the cultural differences between the underlying language communities. Contributing to Wikipedia means more than writing encyclopedic content: it allows communities to store cultural memories of events [22–24], document their point of view [20, 21], and give prominence to people [25]. This collective sifting of culturally-relevant knowledge is such an important social process that conflicts and edit wars frequently emerge before reaching consensus [26]. Finally, the language communities not yet represented on Wikipedia seek the inclusion as an opportunity to establish and promote their language and culture in the digital realm [27]. There are currently 160 open requests for new Wikipedia language editions in the Wikimedia Incubator [28]. Wikipedia is rich

in cultural material, and all data are recorded and openly available, which makes the encyclopedia an attractive object for research on culturally-mediated behaviour.

Quantifying cultural similarity. Multiple numerical measures have been proposed to assess the degree of cultural similarity, although many of them suffer from practical scalability issues or focus on a narrow aspect of culture. The most often cited measure is known as Hofstede's dimensions of culture, which delineates cultures by national borders [29]. Evidence of national cultural differences has been found in the style of collaborative authoring of Wikipedia articles [30, 31]. West [32] quantifies cultural distance through linguistic distance between languages. Several studies delineated cultures by language, and focused on Wikipedia data. In particular, Laufer and colleagues [33] developed measures of cultural similarity, understanding, and affinity through comparing how food cultures are described by self- and foreign communities. Eom *et al.* [34] applied ranking algorithms to biographical articles and obtained a network of cultural agreement on what historical figures are viewed as important, which includes 24 language points of view. Finally, the value of Wikipedia for such anthropological questions as assessing cultural chauvinism or differences in historical world view between cultures has been discussed in [35]. Cultural differences have also been found in other modalities of online communication and collaboration, on such multilingual platforms as Facebook [36], Twitter [37–39], and YouTube [40].

Although previous research has advanced scientific understanding of cultural similarity, attempts to quantify it, for practical reasons, were mostly limited to comparing a small number of cultures along a selected topical dimension. The literature shows a need to establish a scalable approach to quantifying cultural similarity which allows comparing multiple permutations of language dyads and obtaining a bird's-eye view on global intercultural relationships.

3 Data

There are almost 300 language editions of the encyclopedia, which vary greatly in size. This makes sampling a nontrivial decision: on the one hand, many editions are rather small, and sampling from them would not provide data sufficient for statistical analysis. On the other hand, downloading full data on every language edition over a long period of time would be computationally expensive. As a compromise, we focused the analysis on a sample of 126 largest editions which contained more than 10,000 article pages, as of July 2014 [13].

Sampling procedure. To account for variations in editions' age, number of active contributors, and growth rates, we selected the time frame such that (1) to ensure a sufficient amount of editions existed in the beginning of the observation; (2) to allow enough time for each edition to accumulate concepts. We traced back each edition to its first registered article page, and found out that 110 out of 126 largest editions had been created before 01.01.2005. We excluded 11 editions which appeared later (min, vo, be, new, pms, pnb, bpy, arz, mzn, sah, vec) and those whose language codes could not be mapped to the ISO 639-1 standard (be-x-old, zh-yue, bat-smg, map-bms, zh-min-nan). These remaining 110 editions became the focus of our subsequent analysis which covers the period of 9 years between 01.01.2005 and 31.12.2013.

We sampled from each edition separately, collecting IDs of all article pages created between 2005 and 2013 (excluding other types of pages, redirects, and pages created by

bots). For each ID we also collected the entire editing history in all linked language editions. Thus, each ID corresponds to a concept (the topic of the article regardless of the language), and all interlinked language editions represent various linguistic points of view on the concept. After removing duplicates, our dataset includes 3,066,736 unique concepts and a total of 1,360,647,795 article pages in different languages. The data were collected between 20.12.2015 and 25.01.2016 from Wikimedia servers directly, using the access provided by Wikimedia Tool Labs [41].

One algorithmic limitation of our approach is the fact that we rely on Wikipedia's interlanguage link graph to identify articles on the same concepts in different language editions. This approach has some known issues with the lack of triadic closure and dyadic reciprocity [19]. To ensure that the maximal set of interlanguage links related to a concept is retrieved, we collect all articles with their interlanguage links from each edition separately, removing duplicates afterwards. Thus, all existing interlanguage links are extracted.

4 Extraction of co-editing patterns

In this section, we describe the procedure of extracting cultural similarities from co-editing activity in Wikipedia, and present the network of significant shared interests between 110 language communities. The section begins with summarising our pre-analysis check of whether the language-concept overlap in Wikipedia is random.

4.1 Testing for non-randomness of co-editing patterns

Theoretically, each concept covered in Wikipedia could exist in all 288 language editions of the encyclopedia. This is possible because Wikipedia does not censor topic inclusion depending on the language of edition, and anyone is free to contribute an article on any topic of significance. However in practice, such complete coverage is very rare, and concepts are covered in a limited set of language editions. Is this set of languages random? To answer this question, we analyse matrices of language co-occurrences based on a 6.5% random sample of the data (200,748 concepts).

We construct the matrix of empirical co-occurrences C_{ij}, based on the probability of languages i, j to have an article on the same concept. We also construct a synthetic dataset where we preserve the distribution of languages and the number of concepts, $N = 200,748$, but allow languages to co-occur at random. We use the resulting data to produce the matrix of random co-occurrences C_{ij}^{rand}, and compare it to the matrix of co-occurrences C_{ij}. Our null model corresponds to belief that in Wikipedia each concept has equal chances to be covered by any language, with larger editions sharing concepts more frequently purely because of their size. Comparing two matrices allows us to get a preliminary intuition of the extent to which co-editing patterns are non-random.

We establish that language dyads do not edit articles about the same concept (co-occur) by chance. Large editions share concepts more frequently than expected: although in the data *EN-DE* and *EN-FR* overlap in 45% of cases, only 15% is expected by the null model. To little surprise, the amount of overlap between editions in the data decreases with the size of the editions. One notable exception is the Japanese edition which, despite being among the ten largest Wikipedias, co-occurs with other top editions noticeably less frequently. Similarly, the Uzbek edition, being among the ten smallest in the dataset, shows high concept overlap with large editions. By simply plotting frequencies of co-occurrences, we do not observe any local blocks or clusters, neither among large nor small editions (see Figure A1 in Additional file 1).

These overlap differences are statistically significant, and the null model explains only 1,386 out of 11,990 language pairs (11% of observed data, 95% confidence level). Such low explained variation suggests that concept overlap is not random and cannot be explained only by edition sizes. Instead, there are non-random, possibly cultural processes, that influence which languages cover which concepts on Wikipedia. Having evidence that the data contain a signal, we continue our investigation by performing network analysis.

4.2 Inferring the network of shared interest

We look for the languages that are consistently interested in editing articles on the same topics by comparing the differences between observed and expected co-editing activity on each concept. We give a z-score to every language pair, and compare it to the threshold of significance to filter out insignificant pairs. This logic is demonstrated in Figure 1. The result is a weighted undirected network of languages, where languages are connected based on shared information interest.

We first compute the empirical weight w_{ij}^c of a link between languages i, j which co-edit a concept c:
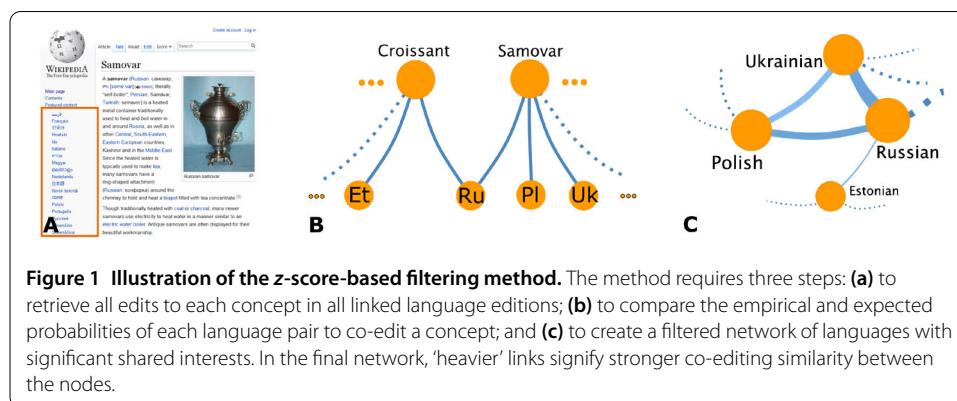
$$w_{ij}^c = k_i^c k_j^c. \tag{1}$$

Here, k_i^c is the number of edits to the concept c in the language edition i, which we use as a proxy to the amount of editing work invested in the concept. This is done across all concepts and language permutations. To determine which links are statistically significant, and which exist purely by chance or due to size effects, we construct a null model where we assume that links between languages i and j are random.

Let the total editing probability of a language be $p_i = \frac{1}{M} \sum_c k_i^c$, where M is the total number of edits for all concepts and language editions. Then the expected probability $\mathrm{E}[w_{ij}^c]$ that languages i and j co-edit the same concept c is:

$$\mathrm{E}[w_{ij}^c] = n_c(n_c - 1)p_i p_j, \tag{2}$$

where n_c is the total number of edits to a concept from all language editions. To compare the difference between observed and expected link weights, we compute a z-score z_{ij}^c for

Figure 1 Illustration of the *z*-score-based filtering method. The method requires three steps: **(a)** to retrieve all edits to each concept in all linked language editions; **(b)** to compare the empirical and expected probabilities of each language pair to co-edit a concept; and **(c)** to create a filtered network of languages with significant shared interests. In the final network, 'heavier' links signify stronger co-editing similarity between the nodes.

each concept and pair of languages i, j, defined as

$$z_{ij}^c = \frac{w_{ij}^c - \mathrm{E}[w_{ij}^c]}{\sigma_{ij}^c}, \tag{3}$$

where σ_{ij}^c is the standard deviation of the expected link weight [42].

Finally, to find the cumulative z-score for a pair of languages i, j, we sum their z-scores over all concepts

$$z_{ij} = \sum_c z_{ij}^c. \tag{4}$$

The relationship between i and j is significant if the cumulative probability of their total z-score, z_{ij} in the right tail falls beyond the p-value $p = 1 - 0.05/N$, where N is the total number of languages. We use the Bonferroni correction [43] to account for the multiple comparisons and size effects in the data. This corresponds to a z-score of 3.32. Since z-scores are sums across many independent variables, their distribution can be approximated by the normal distribution, and the threshold for link significance in the right tail is $t = 3.32\sqrt{L}$, where $L = 3,066,736$ is the number of concepts. We create a link between a pair of languages i, j if the observed z-score, z_{ij}, is above the threshold t [42].

We use the resulting z-scores to build a network of shared topical interests, where the edges are weighted by the similarity of interest, quantifies via z-scores. In summary, this approach allows for discovering significant language pairs of shared interest, accounting for editions of different sizes, and avoiding over-representing the large editions [42].
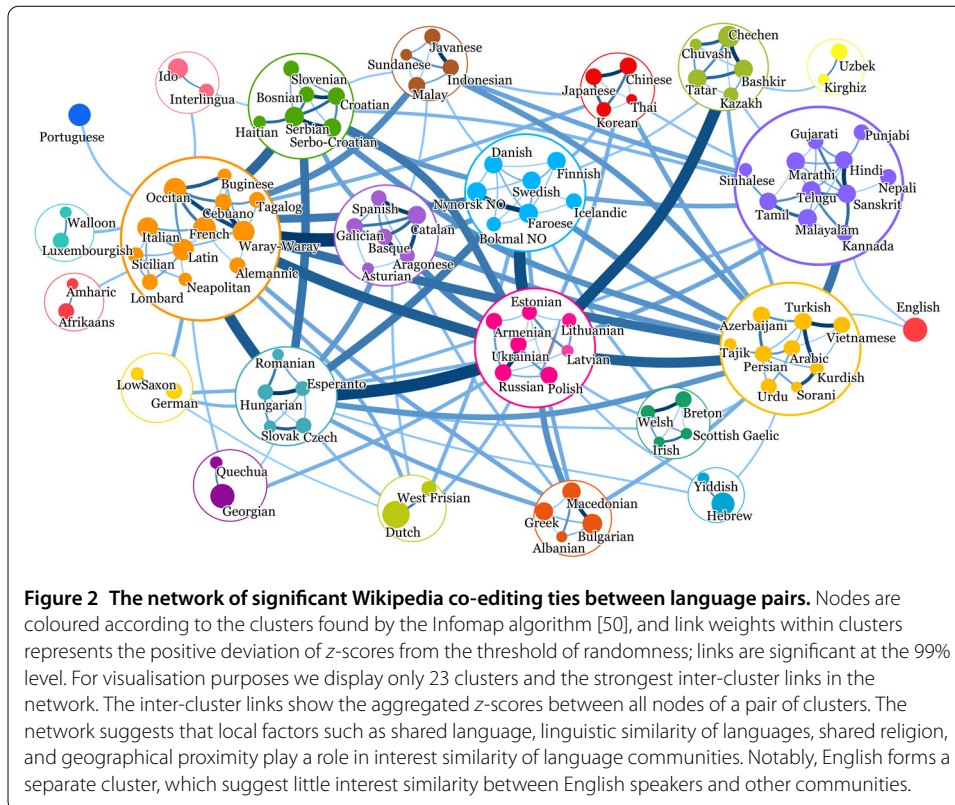
Other methods exist to extract significant weights in graphs. For example, [44] used the hypergeometric distribution for finding the expected link weights for bipartite networks and measured the global p-value. Serrano *et al.* [45] used a disparity filtering method to infer significant weights in networks. Similar to our work, [46] proposed pair-wise connection probability by the configuration model and used the p-value to measure statistical significance of the links.

The network consists of 110 nodes (language editions) and 11,986 undirected edges, and is a complete graph. This means that most languages show at least some similarity in the concepts they edit, however the strength of similarity differs highly across language pairs. The distribution of edge weights is highly skewed with the lowest z-score between Korean and Buginese and the highest z-score between Javanese and Indonesian.

4.3 Clustering the network of significant shared interests

We use the Infomap algorithm [47] to identify language communities that are most similar in their interests. We release a random walker on the network, and allow it to travel across links proportional to their weights. By measuring how long the random walker spends in each part of the network, we are able to identify clusters of languages with strong internal connections [47]. Additionally, we compare these results with the Louvain clustering algorithm [48] and establish that both methods show high agreement.

Our cluster analysis suggests that no language community is completely separated from other communities, and in fact, there are significant topics of common interest between almost any two language pairs. We reveal 21 clusters of two and more languages, plus 9 languages that are identified as separate clusters (see SI for full information on the clusters).

Figure 2 The network of significant Wikipedia co-editing ties between language pairs. Nodes are coloured according to the clusters found by the Infomap algorithm [50], and link weights within clusters represents the positive deviation of *z*-scores from the threshold of randomness; links are significant at the 99% level. For visualisation purposes we display only 23 clusters and the strongest inter-cluster links in the network. The inter-cluster links show the aggregated *z*-scores between all nodes of a pair of clusters. The network suggests that local factors such as shared language, linguistic similarity of languages, shared religion, and geographical proximity play a role in interest similarity of language communities. Notably, English forms a separate cluster, which suggest little interest similarity between English speakers and other communities.

Notably, English forms a self-cluster, and this independent standing means little interest similarity between English and other languages. This is an interesting finding in the light of the recent discussions on whether English is becoming a global language and the most suitable *lingua franca* for cross-national communication [49].

The resulting network is visualised in Figure 2. The links within clusters are weighted according to the amount of positive deviation of *z*-score per language pair from the threshold of randomness. Stronger weights indicate higher similarity. The links are significant at the 99% level. The inter-cluster links should be interpreted with care in the context of this study, as they are weighted according to the aggregated strength of connection between all nodes of both clusters. The network is undirected since it depicts mutual topical interest of both language communities, which is inherently bidirectional. For visualisation purposes, we display only the strongest inter-cluster links and 23 language clusters. Cluster membership information is detailed in Table A1 in Additional file 2.

Cluster interpretation. Visual inspection of language clusters suggests a number of hypotheses which might explain such network configuration. For example, (1) geographical proximity might explain the Swedish-Norwegian-Danish-Faroese-Finnish-Icelandic cluster (light blue), since those are the languages mostly spoken in the Nordic countries. Other groups of languages form around (2) a local *lingua franca*, which is often an official language of a multilingual country, and include other regional languages which are spoken as second- and even third language within the local community. This way, Indonesian and Malay form a cluster with Javanese and Sundanese (brown), which are two largest regional languages of Indonesia. Similarly, one of the largest clusters in the network (purple) consists of 11 languages native to India, where cases of multilingualism are especially common,

since one might need to use different languages for contacts with the state government, with the local community, and at home [49]. Another interesting example is the cluster of languages primarily spoken in the Middle Eastern countries (yellow), which apart from geographical proximity are closely intertwined due to (3) a shared religious tradition. Finally, some clusters illustrate (4) the recent changes in sociopolitical situation, which can also be partially traced through bilingualism. Following the civil war of the 1990s in former Yugoslavia, its former official Serbo-Croatian language is now replaced by three separate languages: Serbian, Croatian, and Bosnian (green cluster). Notably, there is still a separate Serbo-Croatian Wikipedia edition. To give another example, Russian held a privileged position in the former Soviet Union, being the language of the ideology and a priority language to learn at school [49]. Even twenty years after the dissolution of the Soviet Union, Russian remains an important language of exchange between the post-Soviet countries. Similarity of interests between speakers of Russian and the languages spoken in nearby countries, as seen in the magenta cluster, comes as little surprise.
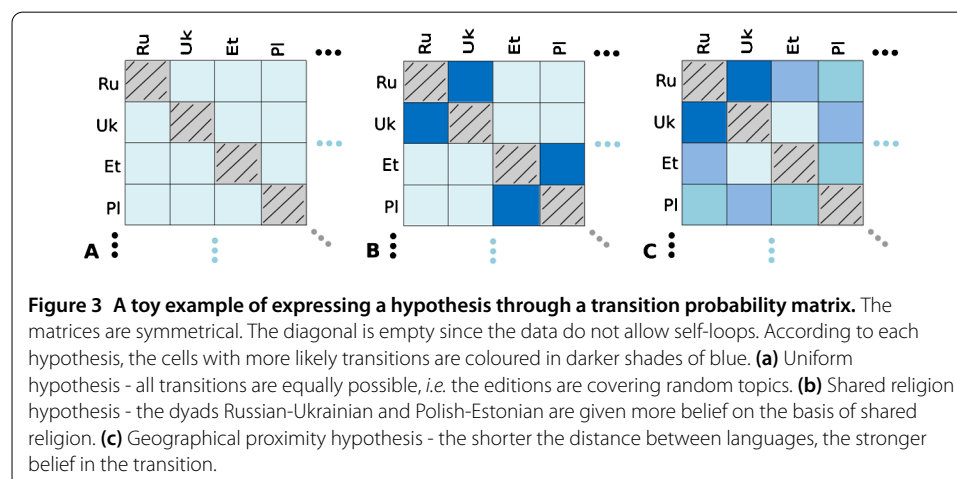
We use this anecdotal interpretation of the clusters to inform our hypotheses about the mechanisms that affect the formation of co-editing similarities. In the next section we will build on these initial interpretations and formulate them as quantifiable hypotheses. To evaluate the validity of the hypotheses, we will compare their plausibility against one another using statistical inference approach.

5 Explanation of co-editing patterns

In this section we show how the network of significant shared interests could be used to inform hypothesis formulation. We compare the plausibility of hypotheses using two statistical approaches. First, we use Bayesian approach and visually compare the strengths of hypotheses. Then we apply frequentist approach to report the explanatory power of different models. We begin by outlining the necessary methodology and continue with reporting the results.

5.1 Hypothesis formulation

We convert our initial interpretation of the network clusters into quantifiable hypotheses, which we express through transition probability matrices illustrated in Figure 3. The hypotheses aim to explain the link weights in the network of co-editing similarities, which

Figure 3 A toy example of expressing a hypothesis through a transition probability matrix. The matrices are symmetrical. The diagonal is empty since the data do not allow self-loops. According to each hypothesis, the cells with more likely transitions are coloured in darker shades of blue. **(a)** Uniform hypothesis - all transitions are equally possible, *i.e.* the editions are covering random topics. **(b)** Shared religion hypothesis - the dyads Russian-Ukrainian and Polish-Estonian are given more belief on the basis of shared religion. **(c)** Geographical proximity hypothesis - the shorter the distance between languages, the stronger belief in the transition.

correspond to the obtained z-scores. The transition probability matrices are square with dimensions $N = 110$, corresponding to the number of language editions studied. The diagonal is empty, since self-loops are not allowed. The formulas, the definitions, and data sources for hypotheses formulation are summarised for reference in Table 1. Below we give more extended explanations on the process of hypotheses construction.

- *H0: Uniform*

    All language co-occurrences are possible with the same probability. A concept can be randomly covered by any language edition. The transition probability t_{ij} for all permutations of languages i and j is

$$t_{ij} = 1.$$

- *H1: Shared language family*

    We retrieve the whole family tree profile of each language and count the number of branches overlapping between each language dyad. For example,
    – Arabic: Afro-Asiatic; Semitic; Central Semitic; Arabic languages; Arabic
    – Hebrew: Afro-Asiatic; Semitic; Central Semitic; Northwest Semitic; Canaanite; Hebrew

    Arabic and Hebrew share three levels of language tree hierarchy (Afro-Asiatic; Semitic; Central Semitic) and thus will have the transition score of 3 in the hypothesis table. If f_i is the set of branches describing the full language family profile of language i, the transition probability t_{ij} corresponds to the count of shared branches in the family tree of languages i and j, and is computed as

$$t_{ij} = |f_i \cup f_j|.$$

- *H2: Bilingual population within a country*

    To formalise other hypotheses, we needed to map languages to countries where they are spoken. We list all countries where a pair of languages are co-spoken; for each country we compute the probability of a person to speak both languages. The hypothesis table contains the average probability of a person to speak both languages computed across all countries where both languages are spoken by more than 0.1% of the population. The transition probability is described by

$$t_{ij} = \frac{1}{N_{ij}} \sum_A p(i)_A p(j)_A,$$

    where $p(i)_A$, $p(j)_A$ are proportions of speakers of languages i, j in a country A, N_{ij} is the number of countries where i, j are co-spoken. The more bilinguals speaking i and j live in the same country, the higher the transition belief.

- *H3: Geographical proximity of language speakers*

    We assign each country with its primary language (the language that the majority of its population speaks) and compute the average distance between all permutations of countries where language i or j are spoken. All inter-country distances are scaled between 0 and 1. Thus,

$$t_{ij} = \frac{1}{N_{ij}} \sum_{A,B} \frac{d_{\min}}{d_{AB}},$$

Table 1 Formalisation of hypotheses to explain the probability of language dyads to co-edit a Wikipedia article about the same concept

Hypothesis and formalisation	Notation	Description	Data source
H0: Uniform hypothesis $t_{ij} = 1$	–	All co-occurrences are equally probable, *i.e.* every edition i covers the same concept as edition j with a constant probability	–
H1: Shared language family $t_{ij} = \|f_i \cup f_j\|$	f_i is the set of branches describing the full language family profile of language i, t_{ij} is the count of shared branches in the family tree of i and j	Language communities of linguistically related languages will show more co-editing similarity	The data on language family classification was taken from English Wikipedia infoboxes of articles on each of 110 languages, such as 'Hebrew language'
H2: Bilingual population within a country $t_{ij} = \frac{1}{N_{ij}} \sum_A p(i)_A p(j)_A$	$p(i)_A$, $p(j)_A$ are proportions of speakers of i, j in a country A, N_{ij} is the number of countries where i, j are co-spoken	Multilingual editors belong to multiple cultural communities and might serve as bridges between them. The more bilinguals speaking i and j live in the same country, the higher the transition belief	Territory-language information was downloaded from [51], and is based on the data from the World Bank, Ethnologue, FactBook, and other sources, including per-country census data
H3: Geographical proximity of languages $t_{ij} = \frac{1}{N_{ij}} \sum_{A,B} \frac{d_{\mathbf{min}}}{d_{AB}}$	N_{ij} is the number of country permutations where i or j are spoken as primary language, d_{AB} is Euclidean distance between each pair of countries, and $d_{\mathbf{min}}$ is the smallest distance between countries in the dataset	The smaller the distance between speakers of i and j living in separate countries, the higher the chances for languages i, j to cover the same concept. We consider one (primary) language per country	Distance between countries is computed as Euclidean distance in kilometers between country capitals [52]
H4: Gravity law - demographic force attracting language communities $t_{ij} = \frac{1}{N_{ij}} \sum_{A,B} \frac{m_{A,i} m_{B,j}}{d_{AB}^2}$	$m_{A,i}$ is the number of speakers of the primary language i in a country A, d_{AB} is Euclidean distance between each pair of counties, N_{ij} is the number of country pairs where i or j are spoken as primary language	The larger the language-speaking population and the smaller the distance between the countries A, B, the more the attraction between i and j. Based on the countries' primary languages	Country population data is taken from CIA Factbook [52]
H5: Shared religion $t_{ij} = \begin{cases} 1, & \text{if } r_i = r_j \\ 0, & \text{otherwise} \end{cases}$	r_i is the dominating religion of a language community. It is defined as the most common religion in the list of countries whose primary language is i	Cultures which profess the same religion will show consistent interest in the same topics	The data on world religions was taken from the most recent 2010 Report on Religious Diversity provided by the Pew Research Center [53]

The hypotheses aim to explain the values of link weights (*z*-scores) in the network of co-editing similarity (see Figure 2 for illustrative purposes). The transition probability matrices are square with dimensions $N = 110$, corresponding to the number of language editions studied. The diagonal is empty, since self-loops are not allowed. The value t_{ij} expresses the hypothesised probability of Wikipedia language editions i and j to cover the same concept. After construction of the hypotheses matrices, the matrices undergo Laplacian smoothing of weight 1 (for HypTrails hypotheses testing only), and are further normalised row-wise. The process is illustrated in Figure 3. The results of hypothesis testing are represented in Figure 4 for the HypTrails approach, and in Figure 2 for the MRQAP approach, and are discussed in Sections 5.2 and 5.3 correspondingly.

where N_{ij} is the number of country permutations where i or j are spoken as primary language, d_{AB} is Euclidean distance between each pair of countries, and d_{min} is the smallest distance between countries in the dataset. The smaller the distance between speakers of i and j living in separate countries, the higher the chances for languages i, j to cover the same concept.

- H4: *Gravity law - demographic force attracting language communities*

  Like in the previous example, we allow one (primary) language per country and consider all country permutations where languages i or j are spoken. Demographic attraction is strongest between large population of speakers who live in separate counties which are located closely. Consider the example of France and Germany, where large numbers of French and German speakers correspondingly, live at close distance. We compute average demographic attraction between all permutations of country pairs. We define

$$t_{ij} = \frac{1}{N_{ij}} \sum_{A,B} \frac{m_{A,i} m_{B,j}}{d_{AB}^2},$$

where $m_{A,i}$ is the number of speakers of the primary language i in a country A, d_{AB} is Euclidean distance between each pair of counties (in kilometers), N_{ij} is the number of country pairs where i or j are spoken as primary language. The larger the language-speaking population and the smaller the distance between the countries A, B, the more the attraction between i and j.
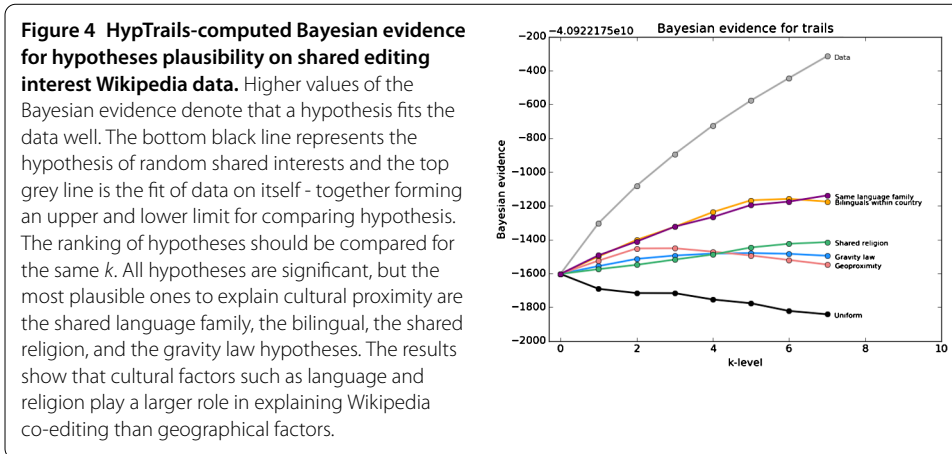
- H5: *Shared primary religion*

  For each country we identified its primary language and its most widespread religion (Christian, Muslim, Hindu, Buddhist, Folk, other or unaffiliated). The religion we assign to a language is the most common religion in the list of countries where the language is spoken as primary. For a language pair, if they share the religion, we add 1 to the hypothesis matrix, and 0 otherwise. Thus the linguistic communities which profess the same religion will show consistent interest in the same topics.

5.2 Bayesian inference - HypTrails

In order to explain why certain languages form communities of shared interest, we need to explain the link weights, or z-score values. We formulate multiple hypotheses based on real-world statistical data, and compare their plausibility using HypTrails [2], a Bayesian approach based on Markov chain processes. We input the z-scores into a matrix, and express hypotheses about their values via Dirichlet priors - matrices of transition probabilities between each possible state (in our case - language edition). We use the trial roulette method to compare different hypotheses. This approach allows to visualise how plausibility of the hypotheses changes with the increasing belief and decreasing allowed variation. Although it was initially designed to compare hypotheses about human trails, in this paper we show that HypTrails is also useful in explaining link weights in networks.

Data preparation. Using the formalisations detailed in Table 1, we fill out corresponding transition probabilities matrices. We apply Laplacian smoothing of weight 1 to all matrices to avoid sparsity issues and to account for the cases when editions co-edit a topic of a general encyclopedic importance which might be relevant for multiple language communities. All matrices are normalised row-wise; diagonals are zero as no self-loops are allowed.

Figure 4 HypTrails-computed Bayesian evidence for hypotheses plausibility on shared editing interest Wikipedia data. Higher values of the Bayesian evidence denote that a hypothesis fits the data well. The bottom black line represents the hypothesis of random shared interests and the top grey line is the fit of data on itself - together forming an upper and lower limit for comparing hypothesis. The ranking of hypotheses should be compared for the same k. All hypotheses are significant, but the most plausible ones to explain cultural proximity are the shared language family, the bilingual, the shared religion, and the gravity law hypotheses. The results show that cultural factors such as language and religion play a larger role in explaining Wikipedia co-editing than geographical factors.

HypTrails ranking. The HypTrails algorithm does not output the absolute values for plausibility of hypotheses, but only compares hypotheses one to another. Thus, one must always compare hypotheses to a uniform hypothesis, and discard those hypotheses that are ranked below the uniform. For the upper bound of comparison, we use the z-scores data itself, since no hypothesis can explain the data better than the data itself.

The results suggest that multiple factors play role in how shared interests are shaped, including geographical proximity, population attraction, shared religion, and especially strongly, linguistic relatedness of the languages and the number of bilingual speakers. No hypothesis explains perfectly all variations in the data, however and all Bayes Factors for all pairs of hypotheses are decisive. Geographical proximity only explains the data to a limited extent, and decays for higher values of k, while the number of bilinguals in the same country, shared language family, and shared religion hypotheses grow stronger with more belief, which suggests that they explain the data most robustly. The explanatory power of hypotheses should be compared for the same values of k, which expresses how strongly we believe in the hypotheses and how much variation is allowed. Figure 4 summarises the results of the HypTrails algorithm. All hypotheses are compared against the uniform hypothesis of random co-occurrence.

5.3 Frequentist approach - MRQAP

In addition to the HypTrails analysis, we use Multiple Regression Quadratic Assignment Procedure (MRQAP) [54] to assess statistical significance of association between the concept co-editing network ties and various hypothesis. This method has a long established tradition in social network analysis as a way to sift out spuriously observed correlations [55], and is well-suited for analysing dyadic data where observations are autocorrelated if they are in the same row or column [3]. We treat the network of concept co-editing as a dependent variable matrix; the independent variable contains the set of hypotheses about the configuration of the network, expressed via hypotheses matrices. Formulation of hypotheses is given in Table 1. We normalise the matrices row-wise in order to standardise the values across matrices. MRQAP is a nonparametric test - it permutes the dependent variables to account for dyadic inter-dependencies. It is also robust against various underlying data distributions [56]. We used 1,000 permutations, which usually suffices for the procedure [57].

MRQAP ranking. The results of the test are in agreement with the hypothesis ranking obtained from applying HypTrails. The number of bilinguals, shared language family, shared religion and demographic attraction are the factors significantly contributing to cultural similarity, as suggested by the t-statistic. By including all five hypotheses into Model 1, we are able to explain 15% of variation in the data. Geographical distance, although a significant factor in several models, is not a very strong one: after excluding the distance hypothesis (Model 2), precision does not decrease. Excluding other hypotheses one by one (Models 3, 4, 5 and 6) lowers precision considerably. Finally, shared language family and bilinguals alone (Models 21 and 22) explain 5% and 7% variation in shared interests correspondingly. The results of the MRQAP are reported in Table 2. Different models include variations of hypotheses combinations that explain the variation in language co-editing ties.

6 Discussion

In this paper, we have used edit co-occurrences data to investigate cultural similarities between language communities on Wikipedia. We have applied a statistical filtering approach to quantify co-editing similarities and build a network of mutual interests. We have utilised the logic of Bayesian and frequentist hypothesis testing to examine what societal features can explain the observed language clusters. Both approaches render similar results, suggesting that cultural proximity and similarity of interests are best explained by bilingualism, linguistic relatedness of languages, shared religion, and demographic attraction of communities. Geographical distance is a weak, and not very significant factor.

Limitations. Our study is not free of limitations, some of which are inherent to the nature of the chosen data. Although we found in the literature mounting evidence that Wikipedia is a promising and rich data source for those interested in mining cultural relations, we agree that it is only one of many possible media where culture might find reflection. Moreover, Wikipedia itself is not free from structural biases, as it reflects the activity of selected technology-savvy, mostly white and male [58, 59], educated, and economically stable social elites. It by no means is representative of the views of general population. However, it is the elites that often drive the cultural, political, and economic processes [1], and thus Wikipedia editors represent a group worthy of being studied. Furthermore, we point out that even though we focus on 110 largest language editions, we still compare the editions at different growth stages and levels of topical saturation. Although this might introduce unforeseen biases, we do not see it as a major limitation, since we focus on aggregated editing activity and only on the articles created between 2005 and 2013. We leave for future research the interesting task of incorporating the time dimension in the analysis and examining how interests shape and change over time.

Additionally, while our approach is quantitative, it requires some subjectivity in interpreting the clusters and formulating hypotheses. To strengthen the internal validity of the study, we inform our reasoning about the hypotheses both in visual analysis of the clusters and in previous literature on the subject. Still, we do not claim to have exhausted all possible hypotheses which could explain the data. Moreover, other formalisations of the selected hypotheses might render different results.

One of the benefits of our approach is that it is free of biases related to topic selection, since we avoid focusing on specific kinds of topics where cultural similarities might be expected. It also scales well in terms of the number of communities and hypotheses that

Table 2 MRQAP decomposition of pairwise correspondence between concept co-occurrence and cultural factors

Model		Bilinguals	Lang. family	Religion	Gravity	Distance[1]	R^2 adj.	*F*-stat.	df	Intercept
1	Estimate	0.0688	0.1074	0.0900	0.0470	−0.0042*	**0.1458**	410.3	11,984	0.0066
	t-statistic	**27.6524**	**23.6158**	**13.4772**	**10.2732**	**−1.3422***				
2	Estimate	0.0676	0.1075	0.0894	0.0464	–	0.1458	512.4	11,985	0.0067
	t-statistic	29.1517	23.6428	13.4200	10.1893	–				
3	Estimate	0.0703	0.1129	0.1022	–	−0.0009*	0.1384	482.3	11,985	0.0067
	t-statistic	28.1932	24.8853	15.4831	–	−0.2989*				
4	Estimate	0.0685	0.1080	–	0.0581	−0.0016*	0.1329	460.5	11,985	0.0074
	t-statistic	27.3119	23.5817	–	12.7773	−0.5225*				
5	Estimate	0.0716	–	0.0916	0.0598	−0.0055*	0.1061	356.9	11,985	0.0075
	t-statistic	28.1697	–	13.4180	12.8396	−1.7256*				
6	Estimate	–	0.1134	0.0881	0.0546	0.0272	0.09140	302.5	11,985	0.0070
	t-statistic	–	24.2095	12.7958	11.5815	9.0453				
7	Estimate	0.0700	0.1129	0.1020	–	–	0.1386	643.1	11,986	0.0067
	t-statistic	30.2487	24.8885	15.5098	–	–				
8	Estimate	0.0703	0.1151	–	–	0.0030*	0.1212	552.2	11,986	0.0076
	t-statistic	27.9237	25.1460	–	–	0.9388*				
9	Estimate	–	–	0.0898	0.0684	0.0272	0.0470	198.2	11,986	0.0079
	t-statistic	–	–	12.7323	14.2619	8.8191				
10	Estimate	0.0700	–	0.0909	0.0590	–	0.1060	474.8	11,986	0.0075
	t-statistic	29.5521	–	13.3370	12.7297	–				
11	Estimate	–	0.1140	–	0.0654	0.0296	0.0790	344.0	11,986	0.0077
	t-statistic	–	24.1755	–	13.9808	9.7791				
12	Estimate	0.0712	0.1151	–	–	–	0.1212	827.8	11,987	0.0076
	t-statistic	30.4703	25.1430	–	–	–				
13	Estimate	0.0738	–	–	–	0.0027	0.0749	486.5	11,987	0.0085
	t-statistic	28.6184	–	–	–	0.8295				
14	Estimate	–	–	–	0.0794	0.0296	0.0342	213.4	11,987	0.0086
	t-statistic	–	–	–	16.7162	9.5508				
15	Estimate	0.0733	–	0.1072	–	–	0.0940	622.8	11,987	0.0076
	t-statistic	30.9368	–	15.9020	–	–				
16	Estimate	–	0.1222	–	–	0.0357	0.0641	411.6	11,987	0.0080
	t-statistic	–	25.9063	–	–	11.8512				
17	Estimate	–	–	0.0936	0.0741	–	0.0409	256.8	11,987	0.0080
	t-statistic	–	–	13.2534	15.5280	–				
18	Estimate	–	–	–	–	0.0372	0.0118	144.1	11,988	0.0090
	t-statistic	–	–	–	–	12.0025				
19	Estimate	–	–	–	0.0861	–	0.0269	333.1	11,988	0.0087
	t-statistic	–	–	–	18.2514	–				
20	Estimate	–	–	0.1144	–	–	0.0217	267.1	11,988	0.0081
	t-statistic	–	–	16.3447	–	–				
21	Estimate	–	0.1233	–	–	–	0.0532	674.9	11,988	0.0081
	t-statistic	–	25.9798	–	–	–				
22	Estimate	0.0746	–	–	–	–	0.0749	972.2	11,988	0.0085
	t-statistic	31.1808	–	–	–	–				

[1]Primary language.

The combination of all hypotheses explains most of the variation in the data (15%). The most plausible explanations are the number of bilinguals and shared religion. The results of MRQAP agree with the ranking of hypotheses by the HypTrails algorithm. All statistics except those labelled with * are significant at the 0.05 level.

could be analysed. In case of research on multilingual data, an important benefit of our approach is that it only uses metadata on user interactions, and understanding the language itself is not required. Finally, it is applicable for any example of collaborative production of a common good where individual activity of participants is recorded.

Discussion of results. Culture is a very complex concept without a definition that is unanimously accepted by Anthropologists, Social Scientists, or Linguists. Although it is universally agreed that cultural communities exist, their borders are very fuzzy and depend on how the researcher defines the term 'culture'. In this work, we focus on the relation between language and culture, and particularly, on how online linguistic expressions can help distil cultural similarities between multilingual communities of Wikipedia editors. An inseparable part of culture, language is only one way of cultural expression, and more studies are needed to explore how other aspects of culture manifest themselves in off- and online world.

Our analysis shows that the decision to write or not to write an article on a certain topic is not a random one. Similar to the idea of national cultural repertoires in the traditional Cultural Sociology [60], we find that various linguistic communities apply different grammars of worth and criteria of evaluation when selecting the topics to cover, that would appeal to the common interest of the language community. Thus, each language edition represents a community of shared understanding with unique linguistic point of view [19–21], its own controversial topics [26], and concept coverage [18].

We demonstrate that similarity of co-editing interests between language communities can be partially explained by the number of bilinguals and by linguistic similarity of the languages themselves. This comes as little surprise, since language is a fundamental part of identity, self-recognition, and culture [5, 11, 61, 62]. It is hard to separate the effects of the number of bilinguals and shared language family from one another, since both might be related: shared vocabulary and grammatical features of the languages from the same language family might explain higher level of bilingualism for these language dyads. Moreover, language choice and bilingualism are an effect of factors galore, such as post-colonial history, education, language and human right policies, free travel, and migration due to political instability, poverty, religious persecutions or work [63, 64]. Finally, cultural similarity defined through Hofstede's four dimensions of values [29] has also been found to relate to language [30, 32].

Shared religion is another uniting factor for language communities. Our finding is in line with Huntington's thesis which argues that cultural and religious identities of people form the primary source of potential conflict in the post-Cold War era [65]. The studies of email and Twitter communication [66] and similarity in country information interests [42] also reveal the patterns that echo religious 'fault lines'.

Population attraction and geographical proximity are the uniting factors that have been extensively discussed in the literature, most relevantly in the context of mobile communication flows [67] and migration [68]. Similar to our results, several studies report gravity laws in online settings, including [69] and [42]. Not only choice of topics to edit, but also online trade in taste-dependent products is affected by distance. For example, [70] finds that proximate countries show more similarity in taste. Notably, this effect only holds for culture-related products such as music. This further supports our finding that there is a relationship between geographical distance and culture, and allows us to speculate that the Internet fails to defy the law of gravity.

The question of whether English is becoming the world's *lingua franca* is an intriguing one [49]. Its central, influential position in the global language network has been reported in networks of book translations, multilingual Twitter users, and Wikipedia editors [1, 71, 72]. On the one hand, such high visibility allows information to radiate between the more connected languages. On the other hand, our study shows that global language centrality plays a minor role in shared interests. Moreover, we show that the domination of English disappears in the network of co-editing similarities, and instead local interconnections come to the forefront, rooting in shared language, similar linguistic characteristics, religion, and demographic proximity. A similar effect has been observed in international markets, where economic competitiveness is linked to the ability to speak a local *lingua franca*, rather than English [73].

7 Conclusions and implications

Out of almost 300 Wikipedia's language editions, 76% have less than 100 active users [13]. Linguistically, this means that those languages are in danger of extinction [64], at least in the online space [74]. Nevertheless, [27] emphasises the role of Wikipedia in helping peripheral languages cross the digital divide, acquire digital functions and prestige as their speakers go online. At the same time, Pentzold [24] describes Wikipedia as a global cultural memory place, access to which depends on the language skills. In his view, Wikipedia is not a mere encyclopedia where facts are documented, but rather a space where the entire collective memories of important events are constructed during a discursive, social process. We show that the topics that each language edition documents are not selected randomly, however small the underlying community of editors. These non-random processes might relate to the fact that each Wikipedia language edition presents a cultural memory place, where the linguistic point of view and the memorable events of that community are negotiated.

Our findings bring some important policy questions for the Wikimedia Foundation, such as: What are the cultural implications of populating editions with automatically translated concepts present in other language editions? Should English Wikipedia aim at becoming an all-inclusive collection of information from other language editions? Should the decision on who and what will be remembered belong to the community of editors, however small, or to an automated algorithm? We hope that our research will inspire dialogue on how similarities between language communities can be used to improve participation of editors speaking peripheral languages and expand the content of smaller editions.

In addition, Wikipedia has a power to mobilise cultural communities around a very important collective task - selecting and archiving important knowledge for future generations. Our analysis sheds light on how cultural similarities are reflected in this process. We also demonstrate that global cultural interconnections are not dominated by one powerful player, but instead follow the locally established 'fault lines' of bilingualism, shared religion and population attraction. We hope that these results will be useful for managers, economists and politicians working in multicultural settings, enthusiastic Wikipedians, academics wishing to study culture via the web, as well as for the public curious about global, intercultural relationships.

Additional material

Additional file 1: Figure A1 contains the heatmap of editing co-occurrence comparison between empirical and experimental data based on a 6.5% sample of the data (*N* = 200,748 concepts) (pdf)
Additional file 2: Table A1 contains the clusters of languages with shared interest as found by the Infomap clustering algorithm (pdf)

Author details
[1]GESIS - Leibniz-Institute for the Social Sciences, 6-8 Unter Sachsenhausen, Cologne, 50667, Germany. [2]University of Koblenz-Landau, Koblenz, Germany. [3]Integrated Science Lab, Department of Physics, Umeå University, Umeå, Sweden.

References
1. Ronen S, Gonçalves B, Hu KZ, Vespignani A, Pinker S, Hidalgo CA (2014) Links that speak: the global language network and its association with global fame. Proc Natl Acad Sci USA. doi:10.1073/pnas.1410931111
2. Singer P, Helic D, Hotho A, Strohmaier M (2015) HypTrails: a Bayesian approach for comparing hypotheses about human trails. In: Proceedings of the 24th international conference on world wide web. WWW2015. ACM, New York
3. Krackardt D (1987) QAP partialling as a test of spuriousness. Soc Netw 9(2):171-186
4. Kroeber AL, Kluckhohn C (1952) Culture: a critical review of concepts and definitions. Peabody Museum of American Archeology and Ethnology, Cambridge
5. Bloomfield L (1945) About foreign language teaching. Yale Rev 34:625-641
6. Hoijer H (1948) Linguistic and cultural change. Language 24:335-345
7. Silvia-Fuenzalida I (1949) Ethnolinguistics and the study of culture. Am Anthropol 51(3):446-456
8. Voegelin CF, Harris ZS (1945) Linguistics in ethnology. Southwest J Anthropol 1:455-465
9. Bucholtz M, Hall K (2008) All of the above: new coalitions in sociocultural linguistics. J Socioling 12(4):401-431. doi:10.1111/j.1467-9841.2008.00382.x
10. Geertz C (1973) The interpretation of cultures: selected essays. Basic Books, New York
11. Kramsch C (1998) Language and culture. Oxford University Press, Oxford
12. Schich M, Song C, Ahn Y, Mirsky A, Martino M, Barabási A, Helbing D (2014) A network framework of cultural history. Science 345(6196):558-562. doi:10.1126/science.1240064
13. Wikipedia (2015) List of Wikipedias. http://en.wikipedia.org/wiki/List_of_Wikipedias. Accessed 24 Sept 2015
14. Petzold T (2011) The uses of multilingualism in digital culture: the case of inter-language linking. PhD thesis, Queensland University of Technology
15. Wikipedia (2015) Notability. http://en.wikipedia.org/wiki/Wikipedia:Notability. Accessed 24 Sept 2015
16. Hecht B, Gergle D (2010) 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. CHI '10. ACM, New York, pp 291-300. doi:10.1145/1753326.1753370
17. Hecht B, Gergle D (2009) Measuring self-focus bias in community-maintained knowledge repositories. In: Proceedings of the fourth international conference on communities and technologies. ACM, New York, pp 11-20. doi:10.1145/1556460.1556463
18. Callahan ES, Herring SC (2011) Cultural bias in Wikipedia content on famous persons. J Am Soc Inf Sci Technol 62:1899-1915. doi:10.1002/asi.21577
19. Bao P, Hecht B, Carton S, Quaderi M, Horn M, Gergle D (2012) Omnipedia: bridging the Wikipedia language gap. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI '12. ACM, New York, pp 1075-1084. doi:10.1145/2207676.2208553
20. Massa P, Scrinzi F (2012) Manypedia: comparing language points of view of Wikipedia communities. In: Proceedings of the 8th international symposium on Wikis and open collaboration. WikiSym '12. ACM, New York. doi:10.1145/2462932.2462960
21. Massa P, Scrinzi F (2011) Exploring linguistic points of view of Wikipedia. In: Proceedings of the 7th international symposium on Wikis and open collaboration. WikiSym '11. ACM, New York, pp 213-214. doi:10.1145/2038558.2038599
22. Keegan B, Gergle D, Contractor N (2011) Hot off the Wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes. In: Proceedings of the 7th international symposium on Wikis and open collaboration. WikiSym '11. ACM, New York, pp 105-113. doi:10.1145/2038558.2038577
23. Keegan BC (2013) A history of newswork on Wikipedia. In: Proceedings of the 9th international symposium on open collaboration. WikiSym '13. ACM, New York. doi:10.1145/2491055.2491062

24. Pentzold C (2009) Fixing the floating gap: the online encyclopaedia Wikipedia as a global memory place. Mem Stud 2(2):255-272. doi:10.1177/1750698008102055
25. Samoilenko A, Yasseri T (2014) The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. EPJ Data Sci 3:1
26. Yasseri T, Spoerri A, Graham M, Kertész J (2014) The most controversial topics in Wikipedia: a multilingual and geographical analysis. In: Fichman P, Hara N (eds) Global Wikipedia: international and cross-cultural issues in online collaboration. Rowman & Littlefield Publishers, Inc., Lanham, pp 25-48
27. Kornai A (2013) Digital language death. PLoS ONE 8:52-64. doi:10.1371/journal.pone.0077056
28. The Wikimedia Incubator (2015) Requests for new languages. https://meta.wikimedia.org/wiki/Requests_for_new_languages. Accessed 24 Sept 2015
29. Hofstede G (1980) Culture's consequences: international differences in work-related values. Sage Publications, London
30. Pfeil U, Zaphiris P, Ang CS (2006) Cultural differences in collaborative authoring of Wikipedia. J Comput-Mediat Commun 12(1):88-113. doi:10.1111/j.1083-6101.2006.00316.x
31. Hara N, Shachaf P, Hew KF (2010) Cross-cultural analysis of the Wikipedia community. J Am Soc Inf Sci Technol 61(10):2097-2108. doi:10.1002/asi.21373
32. West J, Graham JL (2004) A linguistic-based measure of cultural distance and its relationship to managerial values. Manag Int Rev 44:239-260
33. Laufer P, Wagner C, Flöck F, Strohmaier M (2014) Mining cross-cultural relations from Wikipedia - a study of 31 European food cultures. arXiv:1411.4484
34. Eom Y, Aragón P, Laniado D, Kaltenbrunner A, Vigna S, Shepelyansky DL (2015) Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. PLoS ONE. doi:10.1371/journal.pone.0114825
35. Gloor P, De Boer P, Lo W, Wagner S, Nemoto K, Fuehres H (2015) Cultural anthropology through the lens of Wikipedia - a comparison of historical leadership networks in the English, Chinese, and Japanese Wikipedia. In: Proceedings of the 5th international conference on collaborative innovation networks. COINs15
36. Barnett GA, Benefield GA (2015) Predicting international Facebook ties through cultural homophily and other factors. New Media Soc. doi:10.1177/1461444815604421
37. Eleta I, Golbeck J (2014) Multilingual use of Twitter: social networks at the language frontier. Comput Hum Behav 41:424-432
38. García-Gavilanes R, Mejova Y, Quercia D (2014) Twitter ain't without frontiers: economic, social, and cultural boundaries in international communication. In: Proceedings of the 17th ACM conference on computer supported cooperative work & social computing. CSCW '14. ACM, New York, pp 1511-1522. doi:10.1145/2531602.2531725
39. Mocanu D, Baronchelli A, Perra N, Gonçalves B, Zhang Q, Vespignani A (2013) The Twitter of Babel: mapping world languages through microblogging platforms. PLoS ONE. doi:10.1371/journal.pone.0061981
40. Platt E, Bhargava R, Zuckerman E (2015) The international affiliation network of YouTube trends
41. Wikimedia (2015) Wikimedia Tool Labs. https://wikitech.wikimedia.org/wiki/Main_Page. Accessed 24 Sept 2015
42. Karimi F, Bohlin L, Samoilenko A, Rosvall M, Lancichinetti A (2015) Mapping bilateral information interests using the activity of Wikipedia editors. Palgrave Commun 1:15041. doi:10.1057/palcomms.2015.41
43. Dunn OJ (1961) Multiple comparisons among means. J Am Stat Assoc 56(293):52-64
44. Tumminello M, Miccichè S, Lillo F, Piilo J, Mantegna RN (2011) Statistically validated networks in bipartite complex systems. PLoS ONE 6(3):e17994. doi:10.1371/journal.pone.0017994
45. Serrano MÁ, Boguñá M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci USA 106(16):6483-6488. doi:10.1073/pnas.0808904106
46. Dianati N (2016) Unwinding the hairball graph: pruning algorithms for weighted complex networks. Phys Rev E 93:012304. doi:10.1103/PhysRevE.93.012304
47. Rosvall M, Axelsson D, Bergstrom CT (2010) The map equation. Eur Phys J Spec Top 178(1):13-23
48. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
49. Crystal D (2003) English as a global language, 2nd edn. Cambridge University Press, Cambridge
50. Edler D, Rosvall M (2013) The MapEquation software package. http://www.mapequation.org. Accessed 24 Sept 2015
51. (2015) Territory-language information. CLDR charts - unicode common locale data repository. http://www.unicode.org/cldr/charts/latest/supplemental/territory_language_information.html. Accessed 24 Sept 2015
52. CIA (2015) The world factbook. https://www.cia.gov/library/publications/the-world-factbook/. Accessed 24 Sept 2015
53. Pew Research Center (2010) Religious diversity index scores by country. http://www.pewforum.org/2014/04/04/religious-diversity-index-scores-by-country/. Accessed 24 Sept 2015
54. Hubert L, Schultz J (1976) Quadratic assignment as a general data analysis strategy. Br J Math Stat Psychol 29(2):190-241. doi:10.1111/j.2044-8317.1976.tb00714.x
55. Dekker D, Krackhardt D, Snijders T (2003) Multicollinearity robust QAP for multiple regression. In: 1st annual conference of the North American Association for Computational Social and Organizational Science, pp 22-25
56. Dekker D, Krackhardt D, Snijders TA (2007) Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. Psychometrika 72(4):563-581
57. Jackson DA, Somers KM (1989) Are probability estimates from the permutation model of Mantel's test stable? Can J Zool 67(3):766-769
58. Hill BM, Shaw A (2013) The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation. PLoS ONE 8(6):e65782. doi:10.1371/journal.pone.0065782
59. Antin J, Yee R, Cheshire C, Nov O (2011) Gender differences in Wikipedia editing. In: Proceedings of the 7th international symposium on Wikis and open collaboration. WikiSym '11. ACM, New York, pp 11-14. doi:10.1145/2038558.2038561
60. Lamont M, Thévenot L (eds) (2000) Rethinking comparative cultural sociology. Repertoires of evaluation in France and the United States. Cambridge University Press, Cambridge

61. Castells M (2011) The power of identity: the information age: economy, society, and culture, vol 2, 2nd edn. Wiley-Blackwell, Oxford
62. Whorf BL (1940) Science and linguistics. Technol Rev 42(6):229-231
63. Rassool N (1998) Postmodernity, cultural pluralism and the nation-state: problems of language rights, human rights, identity and power. Lang Sci 20(1):89-99
64. Crystal D (2000) Language death. Cambridge University Press, Cambridge
65. Huntington SP (1993) The clash of civilizations? Foreign Aff 72(3):22-49
66. State B, Park P, Weber I, Macy M (2015) The mesh of civilizations in the global network of digital communication. PLoS ONE. doi:10.1371/journal.pone.0122543
67. Krings G, Calabrese F, Ratti C, Blondel VD (2009) Urban gravity: a model for inter-city telecommunication flows. J Stat Mech Theory Exp 2009(7):L07003
68. Simini F, Gonzalez MC, Maritan A, Barabasi A (2012) A universal model for mobility and migration patterns. Nature 484:96-100. doi:10.1038/nature10856
69. Backstrom L, Sun E, Marlow C (2010) Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th international conference on world wide web. WWW2010. ACM, New York, pp 61-70. doi:10.1145/1772690.1772698
70. Blum BS, Goldfarb A (2006) Does the Internet defy the law of gravity? J Int Econ 70(2):384-405
71. Hale SA (2014) Multilinguals and Wikipedia editing. In: Proceedings of the 2014 ACM conference on web science. WebSci '14. ACM, New York, pp 99-108. doi:10.1145/2615569.2615684
72. Hale SA (2014) Global connectivity and multilinguals in the Twitter network. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI '14. ACM, New York, pp 833-842. doi:10.1145/2556288.2557203
73. Bel Habib I (2011) Multilingual skills provide export benefits and better access to new emerging markets. Sens-Public
74. Prado D (2012) Language presence in the real world and cyberspace. In: Net.lang: towards a multilingual cyberspace. C&F éditions, Caen, pp 38-51