ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Ban K, Perc M, Levnajić Z. 2017 Robust clustering of languages across Wikipedia growth. *R. Soc. open sci.* **4**: 171217. http://dx.doi.org/10.1098/rsos.171217

Received: 23 August 2017
Accepted: 18 September 2017

Subject Category:
Engineering

Subject Areas:
complexity/electrical engineering

Keywords:
Wikipedia, language, growth dynamics, data analysis, clustering

Author for correspondence:
Matjaž Perc
e-mail: matjaz.perc@uni-mb.si

Robust clustering of languages across Wikipedia growth

Kristina Ban[1], Matjaž Perc[2,3] and Zoran Levnajić[1,4]

[1]Faculty of Information Studies, Ljubljanska cesta 31A, 8000 Novo Mesto, Slovenia
[2]Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, 2000 Maribor, Slovenia
[3]CAMTP—Center for Applied Mathematics and Theoretical Physics, University of Maribor, Mladinska 3, 2000 Maribor, Slovenia
[4]Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

MP, 0000-0002-3087-541X

Wikipedia is the largest existing knowledge repository that is growing on a genuine crowdsourcing support. While the English Wikipedia is the most extensive and the most researched one with over 5 million articles, comparatively little is known about the behaviour and growth of the remaining 283 smaller Wikipedias, the smallest of which, Afar, has only one article. Here, we use a subset of these data, consisting of 14 962 different articles, each of which exists in 26 different languages, from Arabic to Ukrainian. We study the growth of Wikipedias in these languages over a time span of 15 years. We show that, while an average article follows a random path from one language to another, there exist six well-defined clusters of Wikipedias that share common growth patterns. The make-up of these clusters is remarkably robust against the method used for their determination, as we verify via four different clustering methods. Interestingly, the identified Wikipedia clusters have little correlation with language families and groups. Rather, the growth of Wikipedia across different languages is governed by different factors, ranging from similarities in culture to information literacy.

1. Introduction

The fact that we are able to carry the knowledge from previous generations forward gives us evolutionary advantages that no other species on the planet can compete with. In fact, widespread cooperation among unrelated individuals on the one hand, and our language on the other hand, are the two defining features that distinguish us most prominently from other species [1–4]. For millennia, we have been upholding a cumulative culture, which led to an exponential increase in our cultural output [5]. Our ability to pass on knowledge from generation to

THE ROYAL SOCIETY PUBLISHING

generation relies on the evolution of language [6–10] via a set of grammatical rules that allow infinitely many comprehensible formulations [11,12]. In times of unprecedented technological progress and scientific breakthroughs, the amount of information to carry forward is staggering, and it requires information sharing, worldwide collaboration, algorithmic prowess of search engines, as well as selfless efforts of countless volunteers to maintain, categorize and help navigate what we know. Wikipedia [13] is surely the most famous example of what can come of such efforts.

Thankfully, much of what we know has been digitized [14,15], and the deluge of digital data, along with recent advances in the theory and modelling of social systems and networks [16–23], enables quantitative explorations of our culture that were unimaginable even a decade ago. From enhanced disease surveillance [24], human mobility patterns [25,26] and the spreading of misinformation [27,28], to the universality in voting behaviour [29] and emotional blogging [30,31] to name just some examples, there are virtually no limits to innovative data-driven research that lifts the veil on how we share information, how and with whom we communicate, to where we travel and essentially, on how we live our lives.

Wikipedia [13] has itself been subject to much research scrutiny, in terms of the accuracy of content [32–34], which proved to be better than that of traditional encyclopaedias, and also in terms of intellectual interchanges during its history [35], the evolution of its norm network [36], the dynamics of conflicts and edit wars [37], dynamics of general growth [38,39], circadian patterns of editorial activity [40], language complexity [41] and even in terms of its convergence of academics [42]. Indeed, the open access policy of Wikipedia, along with a useful API and few limits on data accessibility, has made it one of the most researched data repositories in history.

Importantly, while initially research on Wikipedia has been focused predominantly on the English language, recently the focus has been shifting also towards other languages [43–46], and in particular to the cross-cultural dimension of the database. In [47], the authors have studied editing behaviours in multilingual Wikipedia, focusing on the engagement, interests and language proficiency in the primary and secondary languages of the editors. Research revealed that the English edition of Wikipedia displays different dynamics from the Spanish and German editions, and that it plays the broker role in bringing together content written by multilinguals from many language editions. The study also concluded that language remains a formidable hurdle to the spread of content. Similarly, in [48], cultural borders on Wikipedia through multilingual co-editing activity have been studied, showing that the domination of the English language disappears in the network of co-editing similarities, and that instead local connections come to the forefront. An approach has also been proposed there that allows the extraction of significant cultural borders based on the editing activity of Wikipedia users. Most recently, the early adhesion of structural inequality in the formation of Wikipedia has been studied [49].

Here, we use a relatively small subset of Wikipedia, in particular, 14 962 different articles, but each of which jointly exist in 26 different languages. This gives us the opportunity to study growth patterns of collaborative knowledge across time and across different languages. Essentially, we seek to explore how, given an article that exists in many Wikipedias, this article gets ‘translated’ from one language to another. In particular, does an average article appear in various Wikipedias following a prescribed sequence of languages or not? Can, in this regard, Wikipedias be clustered into language groups with shared growth properties? If yes, can these properties be understood in terms of language families, or in terms of cultural and geographical proximity, or perhaps, in terms of information literacy and policy towards IT education? While we do not arrive at conclusive answers for all these questions, we do show that although an average article follows a random path from one language to another, there are nevertheless robust clusters of languages that share very similar growth rates and statistical properties of the dates of birth of article, and striking similarities in the average time delays between the same articles appearing in two different languages. The languages within the identified clusters have little correlation with language families and groups, making a precise statement with regard to what exactly underpins each cluster difficult to provide.

The continuation of this paper is organized as follows. In §2, we first present the data that we use for our research, while in §3, we present the main results. We conclude in §4 with a discussion and an outlook into the future.

2. Data

Non-English Wikipedias grow by both translating from other Wikipedias (typically English one) and writing articles anew. Regardless of which mechanism is predominant, what we wish to study in this

Table 1. Languages used in our study (each corresponding to one Wikipedia). All 26 languages used in our study (rightmost column) organized into language families and groups (first two columns) according to the standard reference [50]. Italian, German and Russian, for example, are all Indo-European languages, but belong to different language groups. For an easier referral in table 2, a symbol is introduced for each language group. Abbreviations in brackets are used later in figures and conform to the ISO 639 standard.

family	group	languages
Indo-European	Germanic ☆	English (en), German (de), Danish (da), Swedish (sv), Norwegian (no), Dutch (nl)
	Italic ✻	Italian (it), Portuguese (pt), Spanish (es), French (fr), Romanian (ro)
	Slavic ✚	Russian (ru), Polish (pl), Czech (cs), Ukrainian (uk), Bulgarian (bg), Serbian (sr)
	Indo-Iranian ●	Persian (fa)
Uralic	Finno-Ugric ✿	Finnish (fi), Hungarian (hu)
Altaid	Turkic ✺	Turkish (tr)
Afro-Asiatic	Semitic ➙	Arabic (ar)
Sino-Tibetan	Chinese ♦	Chinese (zh)
Korean	❡	Korean (ko)
Austro-Asiatic	❠	Indonesian (id)
Japanese	❤	Japanese (ja)

paper is the spreading dynamics of articles across different language editions of Wikipedia. While we expect that the first language for the vast majority of articles is English, what we seek to elucidate is when and in which sequence those articles appear in other Wikipedias, and if there are any stable patterns in this process.

For this study, we first need a suitable dataset that best allows the study of growth patterns across Wikipedias. Our first idea was to look at the articles that jointly exist in the largest number of Wikipedias. By 'jointly exists', we mean that articles can be found on Wikipedia as different language versions of the same article. We realized this identification by relying on English Wikipedia, so we first identify the English article with the corresponding articles in all other languages and then identify them all as the set of articles in different languages corresponding to each other. However, the topical diversity among such articles turned out to be very narrow (they are mostly articles for largest world countries: article for Russia is, at present, the most translated Wikipedia article). Also, the actual set of languages in which these articles jointly exist turns out to be rather small, which is why we deemed this dataset not suitable for our study.

Next, we looked at the major Wikipedias starting with the English one, and tried to identify an ensemble of articles that jointly exist in all of them. We want to have each considered article present in each considered Wikipedia. But, the more Wikipedias we considered (even only large ones), the smaller was the number of articles that jointly exist in all of them. This is expected since each addition of another Wikipedia reduces the number of articles that jointly exist in all considered Wikipedias. Finally, optimizing between the two, we decided to find the biggest ensemble of articles that jointly exist in the biggest set of major Wikipedias. The best situation was found for 26 Wikipedias (languages) for which there were 14 962 articles that jointly exist in all of them, despite the fact that each selection leaves some important Wikipedias out. We verified that the selected ensemble of articles is topically very diverse, covering almost all domains of knowledge. The chosen 26 languages are reported in table 1 along with their language families and groups marked by symbols. For simplicity, in this table, we also introduce an abbreviation for each language (ISO 639 standard) that we will use later.

We create the dataset by storing the date of birth (DOB) for each article in each Wikipedia. We define DOB as the date on which that article first appeared in that Wikipedia, even if it appeared as a stub (short and incomplete Wikipedia article, often referred to as a stub). We, however, removed the articles that never went beyond the stub stage, since they do not convey much information about Wikipedia growth. We disregard the information about the further growth of these articles (as long as they were not left as stubs). Therefore, our dataset for each among 26 Wikipedias is composed of DOBs for 14 962 articles. Alternatively, it can be seen as an ensemble of 14 962 sequences of DOBs that correspond to 14 962 articles, each sequence containing 26 DOBs corresponding to 26 languages. The DOBs of articles

range from the year 2001 to the year 2016. Thus, the formulated dataset allows for the planned study to be conducted by examining the spreading dynamics of articles from 14 962 sequences of DOBs. The data were extracted and stored automatically using a web crawler designed in programming language Python that was running on a parallel computer. The data were obtained in August 2016.

In the following section, we present our results, which are obtained using standard statistical analysis methods, in particular, standard hierarchical clustering algorithm [51]. The clustering algorithm was implemented using programming language Python, with Euclidean distance playing the role of similarity measure (defined differently for each data analysis approach).

3. Results

We begin by showing in figure 1, trajectories over time of six randomly selected articles from our database, as they appear in language after language. The time of appearance of an article in its first language (first DOB), which is the start of each trajectory, is marked with a circle.

It is impossible to discern a clear sequence in which the languages follow each other in succession. Articles are most likely to first appear in languages such as English or German, but apart from this rather expected observation, it is impossible to predict in which language a particular article will appear next. Also, the evolution of articles starts and finishes at different times. Thus, while there may be predictable statistical regularities in the average growth patterns (as we show later), the path an individual article takes from one language to another is random.

Before turning to average growth patterns, we show in figure 2 scatter plots obtained for six different languages, where the DOB of an article in a particular language is displayed in dependence on its first DOB (in whichever language that is). Each article is represented by a small black dot. If a dot falls exactly on the diagonal, it means that this is the first language for that article. The vertical distance between a dot and the diagonal measures the difference between the DOB in the first language and the DOB in the language considered in each plot.

Looking at specific languages, it can be observed that most of the English articles fall either directly on or very close above the diagonal. This confirms that English is the language in which articles are most likely to appear first. As we move further from figure 2*b* to *c*, to the German and the Dutch, it can be observed that more and more dots fall significantly above the diagonal, thus indicating that for these two languages at least some of the articles appear with a considerable delay with respect to when they have appeared first. In figure 2*d*,*e*,*f*, for the Russian, Czech and Persian, the same trend continues, to the point where barely any article falls on the diagonal, thus indicating that the Persian language was never the 'mother language' of an article in our database. As we will show in what follows, the six displayed languages actually belong to six different language clusters, which share notable statistical similarities in their growth patterns.

3.1. Clustering by cumulative number of articles over time

To study regularities in the average growth patterns over different languages, we show in figure 3 a grey-scale heatmap in which the shade of each block encodes the cumulative number of articles that appeared up to a given year in a particular language, where years are displayed horizontally and languages vertically. Unlike the random trajectories displayed in figure 1, here it can be observed that, on average at least, some Wikipedias grow much in the same way as others.

By determining the dendrogram that links together languages that exhibit similar growth rate [51], we find that the 26 languages can be categorized into six clusters. Namely, looking at the dendogram in figure 3 (dark blue lines in particular), we note that classifying the growth rates into six groups (clusters) makes the clearest distinction among those groups. The vertical ordering of languages in figure 3 comes out of the dendrogram, which is shown on the left margin of the heatmap. From this dendrogram, we obtain the clusters that are reported in the first column of table 2.

While it is perhaps expected that the English language would be in a cluster of its own due to its prominence and widespread use around the world, it is nevertheless surprising to observe that for those clusters that contain more than one language, the languages have very little in common in terms of the language family and group to which they belong (table 1). The simplest explanation for the make-up of the clusters, which would be that language similarity and thus the ease of translating give rise to similar growth rates, does not apply. Other columns in table 2 report clusters obtained via different methods, as we explain in the reminder of this section.

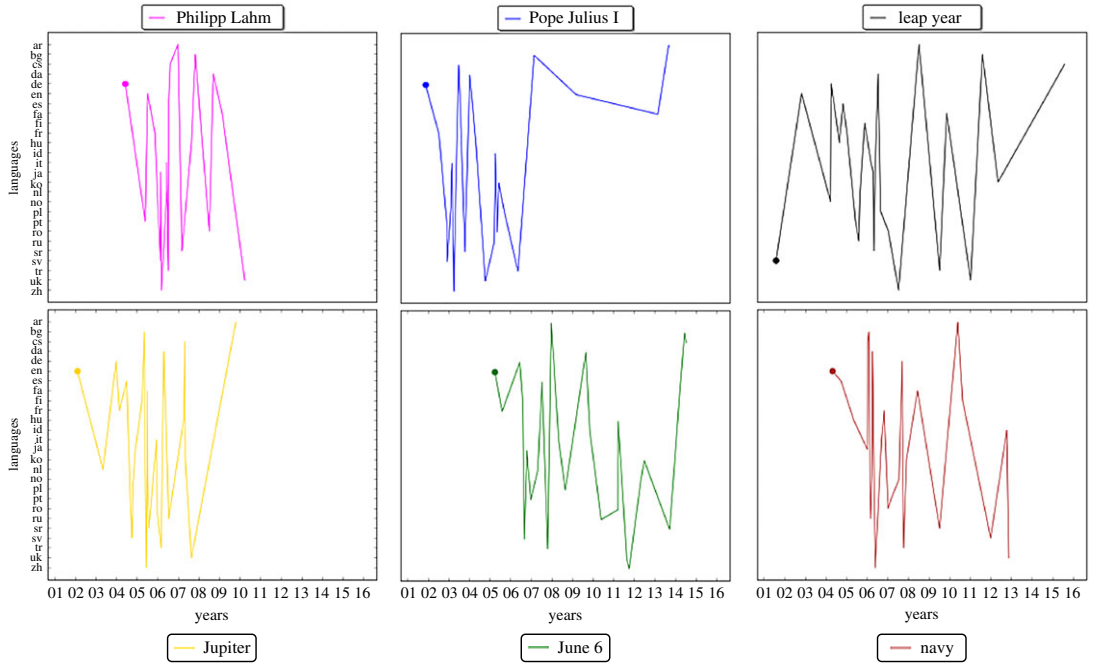

Figure 1. An average article from Wikipedia follows a random path from the language in which first appears to all other 25 languages that our study encompasses. Shown are the trajectories of six randomly selected articles out of the whole database consisting of 14 962 articles. Each trajectory connects the DOBs in all 26 Wikipedias chronologically. The DOB for the first language is marked with a circle for each article. Vertical ordering of languages is alphabetical. Time (denoted by the last two digits of each year) is displayed horizontally. The colour in the trajectories is introduced solely for clarity. The Wikipedia topics of articles are shown in the legend. Presented results are representative in that randomly selecting another subset of articles would give qualitatively the same plot.

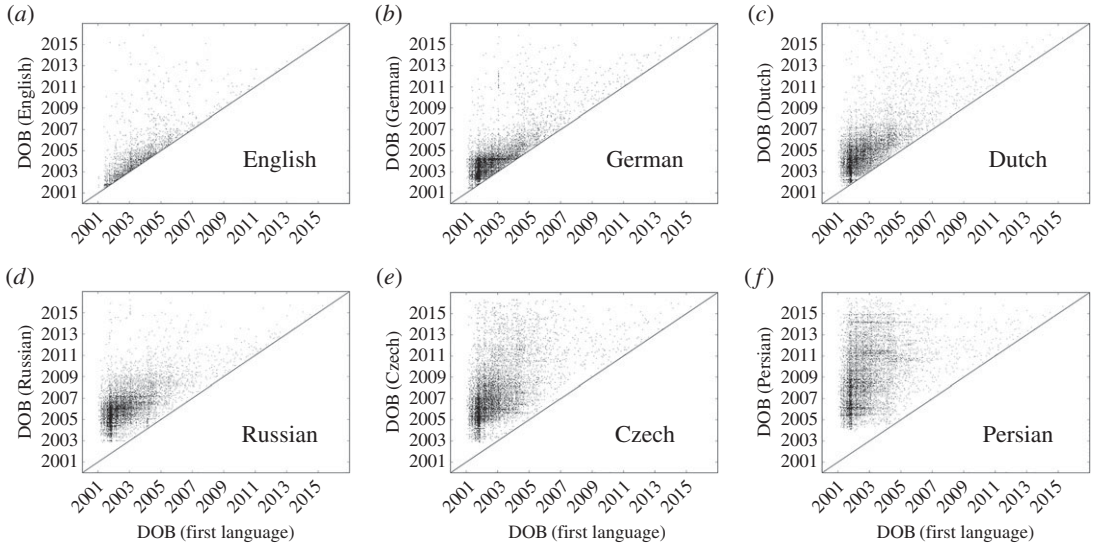

Figure 2. Scatter plots showing the original date of birth (in whichever language) of an article versus the date of birth in the language marked in each plot. Not surprisingly, English is the first Wikipedia for most of the articles, which is reflected in the majority of the dots being on or very close above the diagonal. Moving from (*a*) to (*f*), there is a clear trend of dots scattering further and further away from the diagonal, thus indicating that for these languages the bulk of the articles appear with a growing delay behind their first dates of birth (which is most often in English). The displayed six languages each belong to a different cluster that we discuss in the following figures and in table 2. English and German compose clusters of their own.

3.2. Clustering by average date of birth difference

Of course, determining the clusters based on average growth rates is perhaps not the best, and certainly not the only way to find shared properties among Wikipedias. Arguments could thus be raised whether a

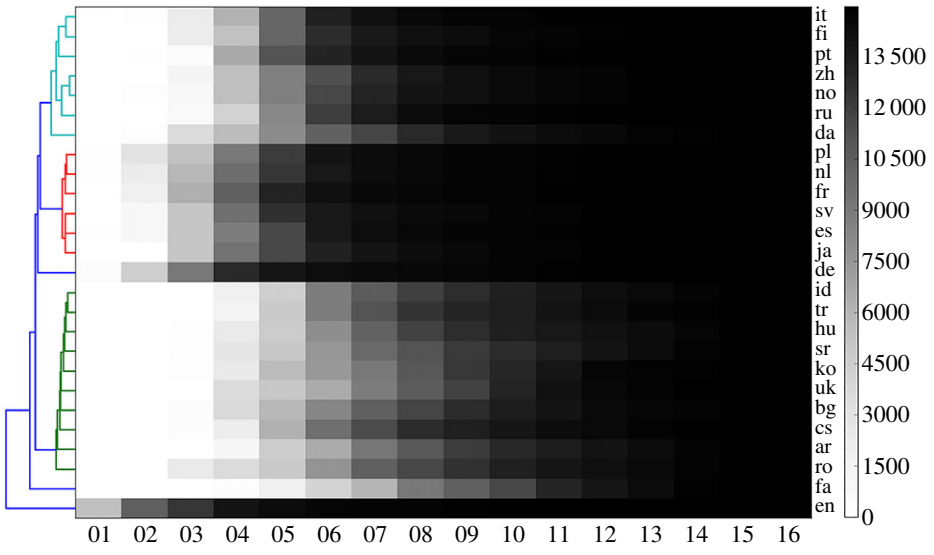

Figure 3. This analysis yields the clusters listed in the first (leftmost) column of table 2. The grey scale in each block encodes the cumulative number (see the bar on the right) of articles that appeared up to that year (horizontally) in that language (vertically). Considering the example of English language, it can be observed that most of the articles considered in our analysis were in existence by the year 2005. This is closely followed by the German articles, then by the French articles, and so on. Notably, this is valid only on average, while each individual article follows a random path, as shown in figure 1. The dendrogram on the left shows the clustering based on the displayed growth rates, such that the languages within each cluster grow similarly fast. Colours in dendrogram indicate the obtained clustering. For the abbreviation of languages, see table 1.

different method of clustering would yield a more expected outcome in terms of grouping of Wikipedias. With this in mind, we now examine a different clustering approach. For each pair of Wikipedias, we considered the difference between DOBs in those two Wikipedias for all articles. Averaging these differences over the entire ensemble of articles, we obtain the heatmap shown in figure 4, where the colour of each block indicates the average DOB difference between the respective pair of languages (Wikipedias). A certain language is on average either ahead (red) or behind (blue) another language, while (almost) white blocks denote that a given pair of languages is practically synchronized, i.e. that articles—on average—appear simultaneously in both Wikipedias.

Using this information as input, we obtain another clustering whose dendrogram is shown on the margins of the heatmap in figure 4. These clusters are listed in the second column of table 2. Although some differences exist in comparison to the clusters determined via previous approach from figure 3, their content is largely the same, and again lacks correlation with the language families and groups listed in table 1. We may thus reiterate the earlier conclusion that the growth of Wikis across different languages is probably governed by a complex interplay of factors beyond languages themselves.

3.3. Clustering from absolute article delays

In contrast to averaging pairwise Wikipedia DOB differences, next we examine the histograms (distributions) of absolute article delays for all Wikipedias. To that end, we consider the difference between article's DOB in a specific language and that article's first DOB (in whichever language it happens to be). We thus obtain delay histograms, which for each Wikipedia capture the distribution of delays of its articles behind their first DOBs. These distributions are shown in figure 5, where each curve is an exponentially modified Gaussian curve, best fitted on the actual histograms (for simplicity and clarity of comparison, we do not show the histograms but only the fitted curves). Clearly, the fact that English is the most common first language is reflected by its distribution being almost perfectly peaked at zero delay. This is true for German to a lesser extent, while the remaining Wikipedias can be neatly grouped into clusters, such that delay distributions within each cluster are remarkably similar in shape, size and peak value of delay. Note that clustering was done in this case using the distances between the curves as the similarity measure, in contrast to previous cases.

Wikipedias in figure 5 are marked by groups of different colours to reflect their organization into clusters that are obtained from this analysis, and are shown in the third column in table 2. Again, we

Table 2. Clusters of languages determined via four different methods. First column (growth): clustering from the individual growth rates over the years (cf. figure 3); second column (DOB): clustering from the averaged differences of DOBs (cf. figure 4); third column (delays): clustering from the statistics of time delays between the first DOB and other DOBs (cf. figure 5); fourth column (MDS): clustering from multi-dimensional scaling of distances between Wikipedia pairs (cf. figure 6). Different clusters are separated by horizontal lines. It can be observed that all 26 languages fall into six different clusters, but the make-up of the clusters changes only slightly depending on the method used. Symbols indicate language groups, as introduced in table 1. Languages within a given cluster (where more than one) are obviously not correlated with a particular language group.

growth	DOB	delays	MDS
English ☆	English ☆	English ☆	English ☆
German ☆	German ☆	German ☆	German ☆
Italian ✲	Italian ✲	Italian ✲	Italian ✲
Finnish ✿	Finnish ✿	Finnish ✿	Finnish ✿
Portuguese ✲	Portuguese ✲	Portuguese ✲	Portuguese ✲
Russian ✚	Russian ✚	Russian ✚	Russian ✚
Norwegian ☆	Norwegian ☆	Norwegian ☆	Norwegian ☆
Chinese ♦	Bulgarian ✚	Chinese ♦	Chinese ♦
Danish ☆	Serbian ✚	Danish ☆	Danish ☆
Polish ✚	Polish ✚	Polish ✚	Polish ✚
Dutch ☆	Dutch ☆	Dutch ☆	Dutch ☆
Spanish ✲	Spanish ✲	Spanish ✲	Spanish ✲
Japanese ❤	Japanese ❤	Japanese ❤	Japanese ❤
French ✲	French ✲	French ✲	French ✲
Swedish ☆	Swedish ☆	Swedish ☆	Swedish ☆
	Danish ☆		
	Chinese ♦		
Indonesian ❣	Indonesian ❣	Indonesian ❣	Indonesian ❣
Turkish ✺	Turkish ✺	Turkish ✺	Turkish ✺
Hungarian ✿	Hungarian ✿	Hungarian ✿	Hungarian ✿
Korean ❡	Korean ❡	Korean ❡	Korean ❡
Ukrainian ✚	Ukrainian ✚	Ukrainian ✚	Ukrainian ✚
Czech ✚	Czech ✚	Czech ✚	Czech ✚
Arabic ➙	Arabic ➙	Arabic ➙	Arabic ➙
Romanian ✲	Romanian ✲	Romanian ✲	Romanian ✲
Bulgarian ✚		Bulgarian ✚	Bulgarian ✚
Serbian ✚		Serbian ✚	Serbian ✚
Persian ●	Persian ●	Persian ●	Persian ●

note striking similarities in the composition of clusters between this method and other two methods employed so far, despite these methods being fairly independent from one another.

3.4. Clustering via multi-dimensional scaling

We conclude this section by presenting yet another clustering approach, conceptually different from the previous ones. We go back to the data behind figure 4, which consists of the average difference between DOBs (value measured as number of days) for each pair of Wikipedias. These values can be understood as pairwise 'distances' between Wikipedias, in the sense that two synchronized Wikipedias will be at distance zero from each other, while a pair of Wikipedias that run ahead/behind each other will be at a certain positive distance from each other, which expresses to what extent they are not synchronized.

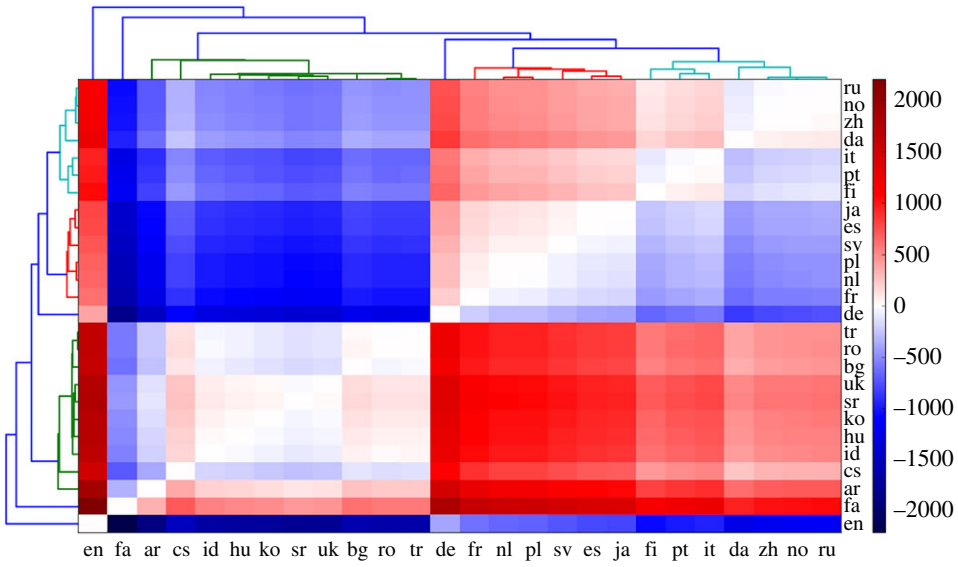

Figure 4. This analysis yields the clusters listed in the second column of table 2. The colour of each block encodes the difference in days between the appearances of articles in two respective languages (see colour bar), averaged over the entire ensemble of articles. The colours in the plot show which Wikipedias are running ahead and which are running behind in a pairwise comparison. White colour indicates that the two languages are practically synchronized, i.e. that articles in those two Wikipedias on average appear simultaneously. It can be observed that the English language is running ahead of all the other languages, while the Persian language is running behind all of them. Also visible are several groups of languages that are well synchronized among them, composing clusters of languages as indicated in the dendrogram on the left and top (the two are identical). Dendrogram is again cut to have six clusters. For the abbreviation of languages, see table 1. Despite the methodological differences, the 26 languages always fall into much the same clusters as observed before in figure 3 (see table 2 for direct comparison).

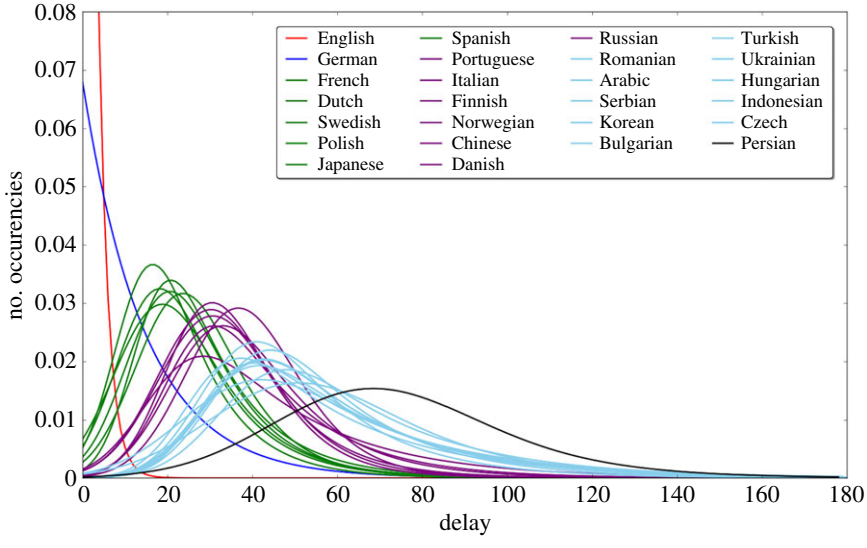

Figure 5. This analysis yields the clusters listed in the third column of table 2. Each curve corresponds to one Wikipedia, and it shows the distribution of delays (in days) of its articles behind the first appearance of these articles (in whichever language the first appearance happens to be). For clarity and simplicity, we do not show the actual data (due to being rather noisy), but we instead show the curves obtained by fitting the actual data to the exponentially modified Gaussian curves, since this way the clustering patterns are more clear. The fits were determined optimally by the shape of the actual data histograms. Different colours are used to distinguish different clusters of languages. Again, English and German are in their separate individual clusters, while the rest of the languages are clustered similarly to what we have already observed in our earlier analysis. The clustering results do not change if we use a different fitting function. This finding further confirms that the composition of the six language clusters is robust and largely independent of the clustering method.

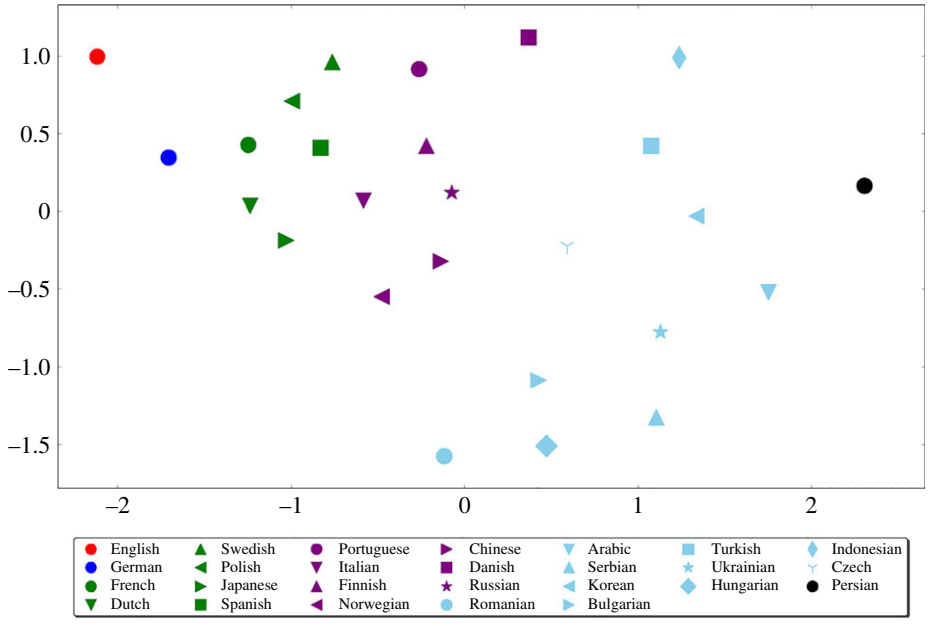

Figure 6. This analysis yields the clusters listed in the fourth column of table 2. Each language (Wikipedia) is represented as a point in two-dimensional space (see legend), while the geometric distance between any pair of points (Wikipedias) captures the averaged difference of DOBs of their articles (values used also in figure 4). Two Wikipedias close together are practically synchronized, while two Wikipedias far apart have one of them running ahead/behind the other. The MDS algorithm visualizes the points in two dimensions while keeping constant the ratios of their respective distances. Clusters of very similar composition can be identified, with English and German again separated from the rest. Each colour indicates languages in one cluster. Note that because of the choice of visualization two-dimensional space, points in some clusters (e.g. cyan) appear further from one another than they really are. This re-confirms the robustness of the obtained clusters to the method used for their determination.

Considering the Wikipedias endowed with pairwise distances among them, we apply the standard multi-dimensional scaling algorithm (MDS) [52] and represent the languages as points embedded in two-dimensional space, as shown in figure 6. MDS transforms the set of pairwise distances keeping the ratio between each two distances into a new set of pairwise distances that can be embedded into a space of given dimensionality. By this procedure, we are able to visualize the 26 Wikipedias as 26 points in space with distances between them illustrating the time delays between each pair of Wikipedias.

We can clearly see the same languages, where English and German are well separated from the rest, each defining its own individual cluster. Remaining languages can be grouped (this time using geometric considerations) into well-localized clusters, each indicated by one colour in figure 6. The clustering coming from this analysis is reported as the last (fourth and rightmost) column in table 2. And yet again, the composition of all six clusters does not substantially differ from the clusterings so far observed, despite a very different approach used this time.

4. Discussion

We have used a relatively small subset of Wikipedia articles that jointly exist in 26 different languages in order to study the commonalities of statistical growth patterns that are shared by Wikipedia editions in different languages. An individual article, in general, follows a random path over time as it is 'translated' from one language to another (we put 'translated' in quotes as we do not maintain that Wikipedias exclusively grow, not only by translating from one to another, but also by writing genuinely new articles). However, upon averaging over the dataset containing 14 962 articles, statistical similarities emerge that quantify how Wikipedias grew in different languages over the last 15 years. In particular, we have observed robust clustering of languages into six distinct clusters, the composition of which is largely independent of the method used for the clustering analysis. This suggests to us that the Wikipedias that share a cluster indeed also share the same overall growth properties, which raises the question as to the reasons behind this. We also verified that considering one or two languages more or less does not have

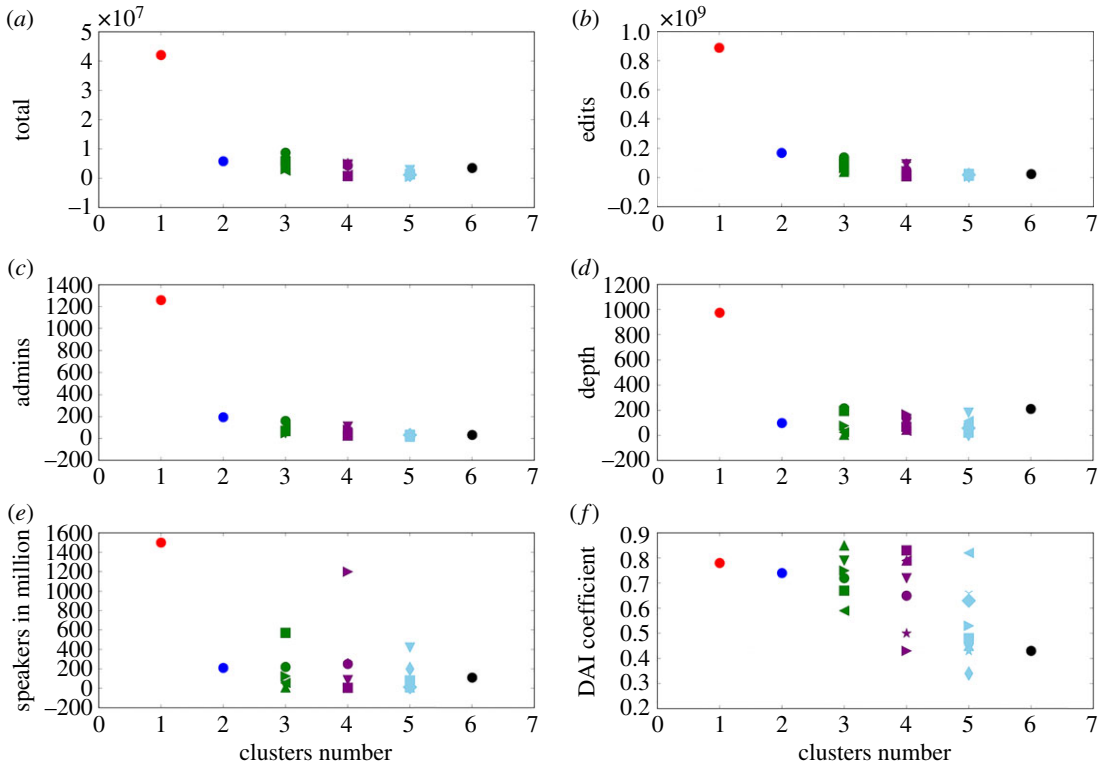

Figure 7. To examine whether the emergence of clusters can be explained by other independent data, we plot six scatter plots where for each language the cluster number (simply from 1 to 6) is on the horizontal axis, while the value for one of the six considered variables is on the vertical axis. The clustering scheme is as in figure 6. Six variables we consider are (*a–f*): total: the total number of Wikipedia pages including both articles and non-articles (e.g. images, talk pages); edits: the number of edits (modifications) in a Wikipedia; Admins: the number of administrators of a Wikipedia; depth: a proxy for the content quality of a Wikipedia defined as (Edits/Articles) × (Non-Articles/Articles) × (1—Stub-ratio); speakers in million: the total number of active speakers of a language; DAI coefficient: Digital Access Index (www.itu.int/ITU-D/ict/dai), a proxy for the IT literacy and accessibility to IT services of a population. All values are obtained independently from the main Wikipedia data. Having the clusters (colours) separated clearly along the vertical axis would amount to a good correlation. However, in (*a*,*b*) and (*c*,*d*) plots (Wikipedia editing parameters), we see a weak correlation that decays towards larger clusters. In (*e*,*f*) (social parameters), we basically find no correlation. This analysis indicates that while some of these variables indeed explain the cluster structure to a modest extent, the complete and conclusive explanation is likely to be hidden in a diverse range of social, economic, IT-related and other factors.

a significant impact on the cluster structure (six clusters can still be identified, but their content depends on which languages we remove from these 26, or which new ones we add on top of these 26).

The simplest explanation revolves around common linguistic or cultural traits that might account for similar 'translatability' of articles. But unexpectedly, our research reveals that for those clusters that contain more than one language, these languages and countries/cultures behind them have very little in common, almost as if the clusters were formed randomly. It is indeed very hard to find any linguistic similarities between Indonesian, Turkish, Hungarian, Korean, Ukrainian, Czech, Arabic and Romanian languages, which are placed in the same cluster by all examined clustering methods. The same is true if we look for possible common cultural traits between populations speaking these languages. This invites the conclusion that the growth of Wikipedia across different languages is governed by a series of factors other than linguistic or cultural familiarity. We arrive at a similar conclusion looking at other multi-language clusters, which is again in favour of a strongly faceted array of factors that together contribute to the formation of the observed clusters.

To test this conclusion more quantitatively, we obtained independent data on six variables that could account for the emergence of clusters. Those are: the total number of Wiki pages, the number of edits in a Wiki, the number of administrators of a Wikipedia, 'depth' of a Wikipedia (proxy for the content quality), the number of active speakers of a language, and the Digital Access Index coefficient (proxy for the IT literacy/accessibility). As we show in figure 7, none of these variables offers a clear and conclusive

explanation of the observed clustering structure, although some variables (e.g. the first three) do show a weak correlation with the clustering structure.

We, therefore, hypothesize that the observed similarities stem from a complex interplay of several social, economic and political factors, probably including at least the following three. First, the total population that regularly communicates in a given language, which roughly overlaps with the joint population of the countries where a given language is in use (while some of the considered languages are spoken in a single country, languages such as English and German are in regular use in several countries). Second, the access to the Internet and average Internet literacy in a country, which has to do with the country's policy towards IT education, which, in turn, correlates with country's level of technological and economic development or even its political order. Third, the general attitude towards the importance of knowledge and education, from which comes the willingness and motivation to volunteer as Wikipedia editor in one's own language, and which, in turn, may depend on how strong are the emotions about 'national culture' in a given country. Since estimating any of these factors is very difficult and the data are not immediately available, we at present cannot offer a conclusive proof of this hypothesis, which we leave as an interesting open question for future work. However, even if we recognize that there are likely to be more relevant factors behind this process, we note that our relatively unexpected findings indicate that the complete set of influencing factors might actually be inferable.

Looking ahead and taking into account very interesting recent developments in research on multilingual Wikipedia [43–49], we are certain that the time is ripe for data-based research on similar databases. This research should be primarily focused on the differences and synergies that emerge as a result of the interactions between different cultures in a worldwide collaborative environment and rely on a large volume of literature on collective social processes [23,53] and spontaneous phenomena such as crowdsourcing [54,55]. Except data-based approaches and statistical modelling, newer IT developments allow for social experiments with several hundreds of participants to be carried out under controlled conditions [56,57]. Our vision is that integrated methodological framework based on data-driven models on the one hand and controlled social experiments on other hand can lead to an interdisciplinary platform for systematic study of collective social phenomena. Constructive synergy between social and computational sciences should here play the crucial role, since the applicative front of social processes is virtually infinite [17,54,56].

Data accessibility. The dataset used for the analysis has been uploaded to the Dryad repository and is available under the http://dx.doi.org/10.5061/dryad.sk0q2 [58].

Authors' contributions. All authors envisaged the research and selected the data and the methods for analysis. K.B. obtained and analysed the data and prepared the figures, M.P. wrote the manuscript, all authors reviewed it.

Competing interests. Authors declare no competing interests.

Funding. This research was supported by the Slovenian Research Agency via projects J1-7009 and V5-1657, programmes P1-0383 and P5-0027, the Young Researcher scheme 36664 and by the EU via MSC-ITN-EJD consortium COSMOS 642563.

Acknowledgements. We thank the Faculty of Information Studies in Novo Mesto for making its computational resources available for this work.

References

1. Miller G. 1981 *Language and speech*. San Francisco, CA: Freeman.
2. Axelrod R. 1984 *The evolution of cooperation*. New York, NY: Basic Books.
3. Nowak MA, Highfield R. 2011 *SuperCooperators: altruism, evolution, and why we need each other to succeed*. New York, NY: Free Press.
4. Hrdy SB. 2011 *Mothers and others: the evolutionary origins of mutual understanding*. Cambridge, MA: Harvard University Press.
5. Lehman HC. 1947 The exponential increase in man's cultural output. *Soc. Forces* **25**, 281–290. (doi:10.1093/sf/25.3.281)
6. Nowak MA, Komarova NL, Niyogi P. 2002 Computational and evolutionary aspects of language. *Nature* **417**, 611–617. (doi:10.1038/nature00771)
7. Abrams D, Strogatz SH. 2003 Modelling the dynamics of language death. *Nature* **424**, 900. (doi:10.1038/424900a)
8. Lieberman E, Michel JB, Jackson J, Tang T, Nowak MA. 2007 Quantifying the evolutionary dynamics of language. *Nature* **449**, 713–716. (doi:10.1038/nature06137)
9. Loreto V, Steels L. 2007 Social dynamics: mergence of language. *Nat. Phys.* **3**, 758–760. (doi:10.1038/nphys770)
10. Solé RV, Corominas-Murtra B, Fortuny J. 2010 Diversity, competition, extinction: the ecophysics of language change. *J. R. Soc. Interface* **7**, 1647–1664. (doi:10.1098/rsif.2010.0110)
11. Chomsky N. 1965 *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
12. Lightfoot D. 1999 *The development of language: acquisition, change and evolution*. Oxford, UK: Blackwell.
13. Wikipedia. 2016 See http://www.wikipedia.org/.
14. Evans JA, Foster JG. 2011 Metaknowledge. *Science* **331**, 721–725. (doi:10.1126/science.1201765)
15. Michel JB, Shen YK, Presser Aiden A, Veres A, Gray MK. 2011 Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182. (doi:10.1126/science.1199644)
16. Albert R, Barabási AL. 2002 Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97. (doi:10.1103/RevModPhys.74.47)
17. Lazer D *et al.* 2009 Life in the network: the coming age of computational social science. *Science* **323**, 721–723. (doi:10.1126/science.1167742)

18. Castellano C, Fortunato S, Loreto V. 2009 Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646. (doi:10.1103/RevModPhys.81.591)
19. Estrada E. 2012 *The structure of complex networks: theory and applications*. Oxford, UK: Oxford University Press.
20. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. 2014 Multilayer networks. *J. Complex Netw.* **2**, 203–271. (doi:10.1093/comnet/cnu016)
21. Boccaletti S, Bianconi G, Criado R, Romance M, Wang Z, Zanin M. 2014 The structure and dynamics of multilayer networks. *Phys. Rep.* **544**, 1–122. (doi:10.1016/j.physrep.2014.07.001)
22. Perc M, Joran JJ, Rand DG, Wang Z, Boccaletti S, Szolnoki A. 2017 Statistical physics of human cooperation. *Phys. Rep.* **687**, 1–51. (doi:10.1016/j.physrep.2017.05.004)
23. Wang Z, Bauch CT, Bhattacharyya S, d'Onofrio A, Manfredi P, Perc M, Perra N, Salathé M, Zhao D. 2016 Statistical physics of vaccination. *Phys. Rep.* **664**, 1–113. (doi:10.1016/j.physrep.2016.10.006)
24. Althouse BM *et al.* 2015 Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Sci.* **4**, 17. (doi:10.1140/epjds/s13688-015-0054-0)
25. González MC, Hidalgo CA, Barabási AL. 2008 Understanding individual human mobility patterns. *Nature* **453**, 779–782. (doi:10.1038/nature06958)
26. Palchykov V, Mitrović M, Jo HH, Saramäki J, Pan RK. 2014 Inferring human mobility using communication patterns. *Sci. Rep.* **4**, 6174. (doi:10.1038/srep06174)
27. Bessi A, Zollo F, Del Vicario M, Scala A, Caldarelli G, Quattrociocchi W. 2015 Trend of narratives in the age of misinformation. *PLoS ONE* **10**, e0134641. (doi:10.1371/journal.pone.0134641)
28. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W. 2016 The spreading of misinformation online. *Proc. Natl Acad. Sci. USA* **113**, 554–559. (doi:10.1073/pnas.1517441113)
29. Chatterjee A, Mitrović M, Fortunato S. 2013 Universality in voting behavior: an empirical analysis. *Sci. Rep.* **3**, 1049. (doi:10.1038/srep01049)
30. Mitrović M, Paltoglou G, Tadić B. 2010 Networks and emotion-driven user communities at popular blogs. *Eur. Phys. J. B* **77**, 597–609. (doi:10.1140/epjb/e2010-00279-x)
31. Mitrović M, Paltoglou G, Tadić B. 2011 Quantitative analysis of bloggers' collective behavior powered by emotions. *J. Stat. Mech.* **2011**, P02005.
32. Giles J. 2005 Internet encyclopaedias go head to head. *Nature* **438**, 900–901. (doi:10.1038/438900a)
33. Chesney T. 2006 An empirical examination of Wikipedia's credibility. *First Monday* **11**. (doi:10.5210/fm.v11i11.1413)
34. Zha Y, Zhou T, Zhou C. 2016 Unfolding large-scale online collaborative human dynamics. *Proc. Natl Acad. Sci. USA* **113**, 14 627–14 632. (doi:10.1073/pnas.1601670113)
35. Yun J, Lee SH, Jeong H. 2016 Intellectual interchanges in the history of the massive online open-editing encyclopedia, Wikipedia. *Phys. Rev. E* **93**, 012307. (doi:10.1103/PhysRevE.93.012307)
36. Heaberlin B, DeDeo S. 2016 The evolution of Wikipedia's norm network. *Future Internet* **8**, 14. (doi:10.3390/fi8020014)
37. Yasseri T, Sumi R, Rung A, Kornai A, Kertész J. 2012 Dynamics of conflicts in Wikipedia. *PLoS ONE* **7**, e38869. (doi:10.1371/journal.pone.0038869)
38. Voss J. 2005 Measuring Wikipedia. In *Proc. 10th Int. Conf. of the International Society for Scientometrics and Informetrics*, *Stockholm, Sweden, 24–28 July* (eds P Ingwersen, B Larsen). Stockholm, Sweden: Karolinska University Press.
39. Suh B, Convertino G, Chi EH, Pirolli P. 2009 The singularity is not near: slowing growth of Wikipedia. In *WikiSym '09: Proc. of the 5th Int. Symp. on Wikis and Open Collaboration*.
40. Yasseri T, Sumi R, Kertész J. 2012 Circadian patterns of Wikipedia editorial activity: a demographic analysis. *PLoS ONE* **7**, e30091. (doi:10.1371/journal.pone.0030091)
41. Yasseri T, Kornai A, Kertész J. 2012 A practical approach to language complexity: a Wikipedia case study. *PLoS ONE* **7**, e48386. (doi:10.1371/journal.pone.0048386)
42. Samoilenko A, Yasseri T. 2014 The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ Data Sci.* **3**, 1. (doi:10.1140/epjds20)
43. Eom YH, Aragon P, Laniado D, Kaltenbrunner A, Vigna S, Shepelyansky D. 2014 Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PLoS ONE* **10**, e0114825. (doi:10.1371/journal.pone.0114825)
44. Yasseri T, Spoerri A, Graham M. 2013 The most controversial topics in Wikipedia: a multilingual and geographical analysis. (http://arxiv.org/abs/1305.5566)
45. Yu AZ, Hu KZ, Jagdish D, Hidalgo CA. 2014 Pantheon: visualizing historical cultural production. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conf., Paris, France, 25–31 October*, pp. 289–290. IEEE. (doi:10.1109/VAST.2014.7042534)
46. Iñiguez G, Török J, Yasseri T, Kaski K, Kertész J. 2014 Modeling social dynamics in a collaborative environment. *EPJ Data Sci.* **3**, 7. (doi:10.1140/epjds/s13688-014-0007-z)
47. Kim S, Park S, Hale SA, Kim S, Byun J, Oh AH. 2016 Understanding editing behaviors in multilingual Wikipedia. *PLoS ONE* **11**, e0155305. (doi:10.1371/journal.pone.0155305)
48. Samoilenko A, Karimi F, Edler D, Kunegis J, Strohmaier M. 2016 Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity. *EPJ Data Sci.* **5**, 9. (doi:10.1140/epjds/s13688-016-0070-8)
49. Yun J, Lee SH, Jeong H. 2016 Early adhesion of structural inequality in the formation of collaborative knowledge, Wikipedia. (http://arxiv.org/abs/1610.06006)
50. Voegelin CF, Voegelin FM. 1977 *Classification and index of the World's languages*. Amsterdam, The Netherlands: Elsevier.
51. Hastie T, Tibshirani R, Friedman J. 2009 *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer.
52. Yaveroglu ON, Malod-Dognin N, Davis D, Levnajic Z, Janjic V, Karapandza R, Stojmirovic A, Przulj N. 2014 Revealing the hidden language of complex networks. *Sci. Rep.* **4**, 4547. (doi:10.1038/srep04547)
53. Szolnoki A, Perc M. 2016 Collective influence in evolutionary social dilemmas. *EPL* **113**, 58004. (doi:10.1209/0295-5075/113/58004)
54. Lee J *et al.* 2013 RNA design rules from a massive open laboratory. *Proc. Natl Acad. Sci. USA* **111**, 2122–2127. (doi:10.1073/pnas.1313039111)
55. Guazzini A, Vilone D, Donati C, Nardi A, Levnajic Z. 2015 Modeling crowdsourcing as collective problem solving. *Sci. Rep.* **5**, 16557. (doi:10.1038/srep16557)
56. Grujic J, Fosco C, Araujo L, Cuesta J, Sanchez A. 2010 Social experiments in the mesoscale: Humans playing a spatial prisoner's dilemma. *PLoS ONE* **5**, e13749. (doi:10.1371/journal.pone.0013749)
57. Grujic J, Eke B, Cabrales A, Cuesta J, Sanchez A. 2012 Three is a crowd in iterated prisoner's dilemmas: experimental evidence on reciprocal behavior. *Sci. Rep.* **2**, 638. (doi:10.1038/srep00638)
58. Ban K, Perc M, Levnajić Z. 2017 Data from: Robust clustering of languages across Wikipedia growth. Dryad Digital Repository. (http://dx.doi.org/10.5061/dryad.sk0q2)