Wikipedia, a Social Pedia: Research Challenges and Opportunities: Papers from the 2015 ICWSM Workshop

Unpacking the Structure of Knowledge Diffusion in Wikipedia: Local Biases, Noble Prizes and the Wisdom of Crowds

Pierpaolo Dondio
School of Computing
Dublin Institute of Technology
pierpaolo.dondio@dit.ie

Niccolò Casnici
Department of Clinical and Experimental Sciences, University of Brescia, Italy
n.casnici@unibs.it

Flaminio Squazzoni
Department of Economics and Management, University of Brescia, Italy
flaminio.squazzon@unibs.it

Abstract

This paper investigates the diffusion of around 100,000 articles about literary authors in 52 versions of Wikipedia. We studied how Wiki versions replicate articles of authors belonging to a particular linguistic group and we collected findings about the potential mechanisms governing the replication process and its fairness. Results showed that diffusion of articles follows a power law, governed by strong preferences among versions, with a high number of isolated articles only present in one Wikipedia version. We found that the English Wiki has a prominent role in diffusing knowledge. However, results also showed that other Wikipedia versions were fundamental to building a rich global corpus of knowledge. Classical Greek and Latin authors resulted the most replicated set of entries. We found that geographic proximity and linguistic similarity was pivotal to explaining mutual links between Wikis. Finally, despite the presence of preference mechanisms, we found how the relative importance that each Wikipedia versions assigns to the set of authors of each language is significantly correlated with an expert-based ranking built on the outcome of various international literary awards, including the Nobel Prize. Moreover, we showed how Wikipedia exhibits a strong *Wisdom of Crowds* effect, with the collective opinion of all the Wikipedia versions showing a correlation with the experts higher than any individual Wikipedia version, with a value for Pearson's' *r* of about 0.9.

Introduction

Wikipedia is the online free encyclopaedia collaboratively written by web users, ranking among the top 10 most visited websites. It has versions in about 240 languages, with 122 languages hosting at least 100,000 articles (as of June 2014). According to Wikipedia guidelines, the reliability of the entries is guaranteed by the rule of the *neutral point of view* (NPOV), according to which every article should "represent fairly, proportionately, and, as far as possible, without bias, all of the significant views that have been published by reliable sources on a topic."

In spite of this NPOV policy, a recent and growing body of literature from both academia and the popular media has shown that the Wikipedia content is not exempt from bias. For instance, (Denning *et al.* 2005) identify six classes of risk: (i) *accuracy*, i.e., the reader cannot be sure which information is accurate and which is not, (ii) *motives*, i.e., the reader ignores the reason why contributors have decided to write an entry, (iii) *uncertain expertise*, i.e., the reader has no information about the contributor's qualifications, (iv) *volatility*, i.e., the articles can be modified over time making the content unstable and difficult to use for citations, (v) *coverage*, i.e., the contributions are not part of a careful plan to organize human knowledge, but mainly represent the interests of a self-selected set of contributors, and (vi) *sources*, i.e., articles do not cite reliable sources.

An important source of bias could come from the coverage dynamics. Recent studies have looked at how the coverage of contents varies across the different versions of Wikipedia, depending on the language. (Halavais and Lackaff 2008) compare the distribution of topics on Wikipedia with the distribution of books and some field-specific academic encyclopaedias. Their results show that Wikipedia's topical coverage is mainly driven by the specific interests of the contributors and the reliability of an individual article strongly depends on the macro-area of that article. (Hecht and Gergle 2009) analyze 15 different language editions of Wikipedia by using a social network analysis approach. They show that in many cases, the contributors of a specific Wikipedia version encoded information that was relevant for them and other users belonging to the same Wikipedia community, but that was less relevant for contributors to other versions of Wikipedia. The authors call this effect a "self-focus bias". This is corroborated by (Kolbitsch and Maurer 2006), who examine the textual content of each article. Their results show that the way famous individuals are described in the English and German versions of Wikipedia is different due to the fact that, depending on the language, some famous local personalities are described extensively, while in other Wikipedia versions they were only cursorily portrayed.

(Warncke-Wang 2012) investigate the relationship between the different languages in Wikipedia by generating a similarity metric based on the concept of coverage. Results show that the similarities between the editions of Wikipedia (i) decreased depending on the increasing geographic distance between the countries, (ii) increased when the size of the Wikipedias (i.e. the number of articles). However, a set of only 9 versions of Wikipedia was used and similarity was defined through an undirected measure, i.e., a Jaccard index, rather than a direct one. Furthermore, by only looking at the relative size of the Wikipedia versions and without a directed link between them, it is difficult to infer properties of the process of the diffusion of articles among each version.

(Eom and Shepelyansky 2013) analyze the presence of local bias by generating three ranking algorithms based on the network structure of Wikipedia and focussing on how different versions evaluate famous persons. Their results confirm a large presence of local heroes, but they also identify a restricted set of *global heroes* (i.e., personalities recognized by the majority of Wikipedia versions), which creates a network of entanglements between cultures. Authors used an undirected network and centrality measures such as PageRank.

Our paper aims to look at how topics are replicated across multiple Wikipedias and to study interesting properties of the mutual relationships among Wiki versions. We wanted to measure knowledge diffusion in order to look at the tension between the existence of a dominant global language and the presence of localized languages. Secondly, we wanted to understand if diffusion bias was generated by preference mechanism based on geographical or linguistic factors, which follows the well-known *homophily* argument, i.e. the extent to which communicating individuals are similar depending on some common features (McPherson 2001), (Lazarsfeld et al. 1954). Lastly, we wanted to ascertain the extent to which the relative importance that each Wikipedia version assigns to each set of authors of each language can be compared to an expert-based ranking of world literature based on the outcome of various international literary awards, including the Nobel Prize.

The paper is organized as follows: section 2 illustrates the dataset; section 3 analyses the diffusion mechanisms of articles in Wikipedia; section 4 investigates the fairness of the various Wikipedia versions in replicating articles; and a final section presents our conclusion.

Dataset and Naming Conventions

We selected a sample of Wikipedia articles that included the following Wikipedia categories: *Writers* (including *Novelists*), *Poets* and *Philosophers*, from now on referred as *authors*. These articles are of particular interest in studying knowledge diffusion across Wikipedia. It is likely that the original works of authors were first accessible to people speaking their own language and they were translated abroad when they reached an international status.

Selection and classification of authors was based on the language used by the author and her/his nationality. Therefore, the *English* set included *British, American, Irish* and all authors whose mother tongue was *English*. We relied on Wikipedia's classification system. For instance, we identified the list of Swedish authors according to the list of Swedish authors presented in the Swedish Wikipedia. By doing so, we presumed that each list of authors was more complete in its own linguistic version. Each list of authors includes not only world-famous authors but also local authors included in only a small number of Wikipedia versions.

N	Language	W_ℓ	$\|W_\ell\|$	$\|\mathcal{A}_\ell\|$	N	Language	W_ℓ	$\|W_\ell\|$	$\|\mathcal{A}_\ell\|$
1	English	*en*	4532	14408	27	Turkish	*tr*	229	1662
2	Dutch	*nl*	1779	2361	28	Slovak	*sk*	193	818
3	German	*de*	1726	6436	29	Danish	*da*	188	3650
4	Swedish	*sv*	1626	3266	30	Basque	*eu*	182	319
5	French	*fr*	1514	9242	31	Lithuanian	*lt*	165	973
6	Italian	*it*	1127	5939	32	Bulgarian	*bg*	162	1033
7	Russian	*ru*	1120	3128	33	Hebrew	*he*	158	332
8	Spanish	*es*	1106	7338	34	Croatian	*hr*	146	1272
9	Polish	*pl*	1050	2825	35	Slovenian	*sl*	141	1040
10	Vietnamese	*vi*	929	422	36	Estonian	*et*	124	758
11	Japanese	*ja*	913	5397	37	Armenian	*hy*	121	359
12	Portuguese	*pt*	830	1092	38	Galician	*gl*	114	902
13	Chinese	*zh*	774	1083	39	Hindi	*hi*	112	503
14	Ukrainian	*uk*	507	2142	40	Latin	*la*	108	488
15	Catalan	*ca*	429	539	41	Greek	*el*	102	1044
16	Norwegian	*no*	422	1097	42	Azerbaijani	*az*	101	994
17	Persian	*fa*	394	450	43	Thai	*th*	88	424
18	Finnish	*fi*	348	2338	44	Occitan	*oc*	87	229
19	Indonesian	*id*	343	331	45	Georgian	*ka*	83	356
20	Czech	*cs*	297	2029	46	Belarusian	*be*	73	736
21	Arabic	*ar*	283	1100	47	Latvian	*lv*	55	376
22	Korean	*ko*	279	905	48	Urdu	*ur*	52	385
23	Malay	*ms*	264	102	49	Bosnian	*bs*	51	222
24	Hungarian	*hu*	261	4069	50	Albanian	*sq*	51	686
25	Serbian	*sr*	248	798	51	Afrikaans	*af*	45	347
26	Romanian	*ro*	244	560	52	Icelandic	*is*	38	308

Table 1 *– Wikipedia versions considered in this study.*

Moreover, an author could have written in more than one language and have had more than one nationality. However, the impact of these disputed authors was not statistically significant. There were only 311 disputed articles. In these cases, authors were assigned to both versions of Wikipedia.

We also considered Latin and Greek languages, which are useful in analysing the diffusion of classical studies and the impact of two former *linguae francae*. Latin authors were an exception to the strict language-based classification. Latin authors included authors of Ancient Rome (such as *Cicero*) and authors of the Early Middle Ages (such as *Boethius*), while authors of the late Middle Ages (from the 9th century onwards) who used Latin as their first language were classified according to their nationality. For instance, F. Bacon was an English philosopher and Thomas Aquinas was an Italian one.

Table 1 shows the list of the Wikipedia versions considered here. For each version, the table includes the total number of articles for each version (column $|W_\ell|$, expressed in thousands and used to sort the table), and the number of articles in our dataset (column *Authors*). We selected 99,841 articles from 52 different languages. This distribution indicated that Anglophone authors were the biggest cohort, followed by the other major European languages.

Names, symbols and basic notations

We call W_ℓ the Wikipedia version written in ℓ, and $a_\ell \in W_\ell$ the Wikipedia page written in ℓ about article a . We used the Wikipedia code to identify each language as shown in Table 1. We called $\mathcal{A}$ the set of all the articles in our *dataset of* 99,841 *authors*. We called $\mathcal{A}_\ell$ the set of authors to which we assigned the language ℓ. For instance, $\mathcal{A}_{en}$ includes Shakespeare and Hemingway, $\mathcal{A}_{fr}$ includes Proust, Voltaire and so on. W_{en} is the English Wikipedia, which might contain a version of some (though probably not all) the articles in $\mathcal{A}_{fr}$.

We say that a Wikipedia version W_ℓ written in ℓ *owns* all the articles of the authors associated with its language, i.e. $\mathcal{A}_\ell$, while it *replicates* some of the other authors associated with a foreign language different from ℓ. We call $\mathcal{F}_\ell$ the set of all the articles included in W_ℓ about authors associated with a foreign language. For instance, $\mathcal{F}_{en}$ includes Proust and Dante, while $\mathcal{F}_{fr}$ includes Shakespeare and Dante, but $\mathcal{F}_{fr}$ may not include a local Italian writer X, since the French Wikipedia W_{fr} does not contain an article for X.

Given an article a, we called $N(a)$ the number of Wikipedia versions replicating article a. The number of versions $N(a)$ was an index of the global diffusion of an article. A high value for $N(a)$ meant that a was included in a high number of versions, evidence of a universally accepted topic, while a low value for $N(a)$ was evidence of a local topic.

The Diffusion of Articles in Wikipedia

Article Distribution across Wikipedia versions

In this section we start analysing how the set of authors collected is distributed and replicated among each Wikipedia version. Figure 1 (dark grey line) shows the frequency distribution of the articles in our sample by the number of Wikipedia versions $N(a)$ replicating the article.

To check whether the distribution of articles was uniform among all the Wikipedia versions, we simulated a random distribution in which every article was considered to have the same likelihood of being replicated by other Wikipedia versions, i.e., a situation with no preferences in the replication process. The simulation guarantees that each Wikipedia version had the same number of articles owned and replicated as the observed actual Wikipedia versions, but the article to be replicated was selected according to a uniform distribution. The light grey line in Figure 1 indicates the distribution obtained by a random simulation. If the various Wikipedia versions replicated articles following a uniform random distribution, there would be a majority of articles replicated in 4 to 7 versions, very few isolated articles and virtually no globally replicated articles. However, results showed that the actual distribution follows a power law with an exponent $k \approx 1.9$. This implies the presence of a number of isolated or very local articles higher than random, and a number of globally covered articles also higher than random. Table 2 shows the distribution of articles divided into six groups according to their number of versions. More than 62% of articles were isolated articles, 1.78% had more than 20 versions, in comparison to none in the random case. Once data were weighted according to the number of versions of each article, we found that the articles with more than 20 versions accounted for more than 21% of all the pages (0 in the random case) and the articles with $N(a) \leq 3$ were about 40% against 7.4%.

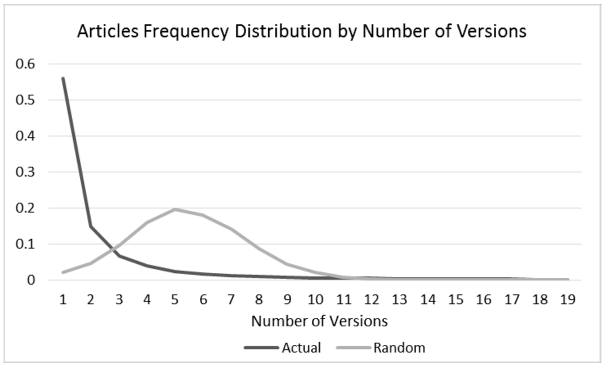

Figure 1 *–Frequency Distribution of Articles.*

Groups **(by value of $N(a)$)**	**N. of Articles**		**N. of Wiki Pages**	
	Actual	**Rand**	**Actual**	**Rand**
Isolated (1)	62.1%	2.04%	22.3%	0.4%
Local (2-3)	21.5%	14.3%	17.8%	7.0%
Regional (4-9)	11.1%	80.6%	22.4%	86.9%
Macro-Region (10-20)	3.5%	3.03%	17.1%	5.78%
International (21-49)	1.7%	0	19.9%	0
Global > 50	0.08%	0	1.47%	0

Table 2. *Frequency distribution of articles by number of versions.*

The presence of a power law indicates that distribution of Wikipedia articles is governed by a preference mechanism where a few articles were highly replicated while a large number of articles were isolated. In other words, the results showed that a restricted minority of articles are common to most Wikipedia versions. This small subset of articles represents the common body of knowledge that can be accessed by individuals speaking most of the languages in our sample. To understand which of the Wikipedia versions was more prone to hosting common knowledge or local articles we calculated the proportion of isolated articles over the total number of articles for each Wikipedia version (Table 3). The table shows whether each of the Wikipedia versions had a proportion of isolated articles significantly higher (symbol '+') or lower (symbol '-') than the average of all the Wikis considered, or whether it did not differ significantly (symbol '='). On the one hand, results showed that English, German and Arabic versions of Wikipedia alone hosted a number of isolated articles consistent with the average. On the other hand, a large group of Wikipedia versions hosted a significantly higher number of isolates, with the Hungarian Wikipedia including about 82% of articles that were not replicated by any other Wikipedia version. A consistent group of Wikipedia versions hosted significantly fewer isolated articles than the average, with the Latin Wikipedia having only 4% of isolates.

The results supplied key indications about how knowledge is distributed into the Wikipedia environment. Wikipedia versions had different features, not only in terms of *quantity* of articles hosted but especially in terms of *type* of article hosted. While many versions of Wikipedia tended to host chiefly local content, other Wikipedia versions played a more crucial role in spreading the knowledge in multiple languages. Consistent with (Hecht and Gergle 2009), the first category of Wikipedia versions encoded information which was just relevant for the users belonging to the same Wikipedia community, focusing mainly on "local heroes" (*ibid.*). Access to the content of these Wikipedia versions is limited to speakers of the local languages and the contributions of such local versions to the common knowledge embodied in the whole encyclopaedia is almost unimportant. In summary, our results are fully consistent with the "self-focus bias" phenomenon by (Hecht and Gergle 2009). On the contrary, the second category of Wikipedia versions provides an important contribution to the spread of knowledge, hosting more common articles than the others. Thanks to the openness of these Wikipedia versions, articles can be read worldwide.

	W	P_i	t		W	P_i	t
1	*hu*	0.82***	+	37	*en*	0.54	=
2	*th*	0.80***	+	38	*fr*	0.54*	-
3	*az*	0.79***	+	41	*es*	0.48**	-
4	*fi*	0.78***	+	42	*el*	0.479***	-
5	*ja*	0.78***	+	43	*pt*	0.473***	-
6	*bg*	0.76***	+	44	*lv*	0.46***	-
7	*sv*	0.75***	+	45	*ca*	0.44***	-
8	*vi*	0.72***	+	46	*he*	0.43***	-
9	*pl*	0.72***	+	47	*fa*	0.43***	-
10	*tr*	0.71***	+	48	*af*	0.41***	-
18	*it*	0.64**	+	49	*ro*	0.4***	-
22	*ru*	0.63**	+	50	*ur*	0.37***	-
24	*zh*	0.62*	+	51	*oc*	0.24***	-
27	*ar*	0.59	=	52	*la*	0.04***	-
35	*de*	0.56	=				

Table 3 *– Proportion P_i of isolated articles. '***' represents 0.99 confidence level, '**' 0.95 and '*' 0.9.*

Measuring preferences among Wiki versions

Up to now our analysis has shown that a vast group of Wikipedia versions replicated more articles than the others and some of these included more global content than others. However, we have no information on the structure of the hosting relationships between versions of Wikipedia, e.g. who is hosting who. The aim of this section is to establish whether Wikipedia could be described as a random chaotic environment where the different versions mutually share contents without any pattern or whether, alternatively, the relationships between versions of the encyclopaedia (or at least some of them) could be explained by the presence of strong preference mechanisms.

First of all, we checked whether the way a Wikipedia version hosts authors of a foreign language significantly deviates from the random expected behaviour.

Given a Wikipedia version W_a identified by its language a and the set of authors $\mathcal{A}_b$ associated with another language b and owned by another Wikipedia version W_b, we calculate a *hosting preference index* $H_z(a,b)$ of W_a with respect to authors $\mathcal{A}_b$. $H_z(a,b)$ is the value of the *t* statistic of the null hypothesis test between the observed number of articles in $\mathcal{A}_b$ replicated by W_a and the expected number of articles.

More specifically, let $T = |\mathcal{A}_{all}|$ be the total number of authors in our dataset (i.e., 99,841). $\mathcal{F}_a$ is the number of articles replicated by Wikipedia a, probably including some versions of articles in $\mathcal{A}_b$. Wikipedia a could host $T - |\mathcal{A}_a|$ articles (all except its own), and if all the articles had the same probability of being hosted, the expected proportion p_e of articles from $\mathcal{A}_b$ hosted by W_a is as follows:

$$p_e(a,b) = \frac{|\mathcal{A}_b|}{T - |\mathcal{A}_a|}$$

However, if the observed number of articles of $\mathcal{A}_b$ hosted by W_a in our dataset is $H(a,b)$, the observed proportion of articles is:

$$p_o(a,b) = \frac{H(a,b)}{\mathcal{F}_a}$$

(number of articles of $\mathcal{A}_b$ hosted by W_a divided by all the foreign articles hosted by W_a). The statistical test compared the two proportions p_o and p_e with $= |\mathcal{F}_a|$. The test standardized score was the value of $H_z(a,b)$. Given a confidence level, the value of $H_z(a,b)$ indicates if the null hypothesis was rejected. If the null hypothesis is accepted, this meant that W_a replicated authors in $\mathcal{A}_b$ as expected, while if the hypothesis is rejected, then the sign of H_z indicates that a hosted higher-than-average or lower-than-average numbers of articles from the set $\mathcal{A}_b$. High positive values indicate a propensity of W_a to accept articles from $\mathcal{A}_b$, while low negative ones indicate that articles from $\mathcal{A}_b$ were hardly replicated by Wikipedia W_a.

If we consider all the pairs among the 52 versions, identifying 2652 possible directed pairs and we set a significance level of 0.99, 1626 pairs (60.9% of the total) had the expected level of hosting, while 413 (16%) were higher than expected and 613 (23.1%) were lower. This indicates that hosting dynamics was not driven by the chance, but on the contrary was often *biased* by strong preferences.

Table 4 shows various measures of the level of preference attributed to each Wikipedia W_ℓ (or better, the set of authors $\mathcal{A}_\ell$ associated with the language of Wikipedia version W_ℓ). For each Wikipedia version, it includes the average value of H_z and the number of Wikipedia versions (out of 51) where each version was hosted significantly more than expected (N_Z^+), less than expected (N_Z^-) and as expected ($N_Z^=$). The difference $N_Z^+ - N_Z^-$ was called Δ_Z and indicates the overall degree of preference attributed to a Wiki. The left-hand side of the table lists the most preferred Wikis, while the right-hand side the less preferred. Results showed the supremacy of classical authors (both *Latin* and *Greek* receive a positive preference from 50 out of 51 Wikipedias and they have the highest value of H_z), followed by English, German, and French. The columns on the right show the bottom 10 Wikipedia versions by average value of H_z. The Hungarian Wikipedia, for instance, had the lowest average preference index and it was hosted less than expected by 49 out of 51 versions.

	$\mathcal{A}_\ell$	$\overline{H_Z}$	Δ_Z , $N_Z^+, N_Z^=, N_Z^-$
1	*la*	19.04	50 (50,1,0)
2	*el*	11.76	50 (50,1,0)
3	*en*	6.29	31 (32,18,1)
4	*de*	4.94	31 (31,0,20)
5	*fr*	3.93	21 (22,28,1)
6	*he*	3.61	23 (23,28,0)
7	*ru*	2.92	14 (20,25,6)
8	*fa*	2.69	18 (18,33,0)
9	*ar*	1.55	14 (15,35,1)
10	*zh*	1.21	6 (8,41,2)

	ℓ	$\overline{H_Z}$	Δ_Z , $N_Z^+, N_Z^=, N_Z^-$
1	*hu*	-7.95	-47 (2,0,49)
2	*ja*	-4.94	-33 (4,10,37)
3	*no*	-4.30	-32 (2,15,34)
4	*id*	-4.23	-26 (0,25,26)
5	*uk*	-4.11	-32 (3,13,35)
6	*da*	-3.9	-41 (2,6,43)
7	*nl*	-3.87	-34 (0,17,34)
8	*es*	-3.84	-35 (5,6,40)
9	*gl*	-3.51	-28 (2,19,30)
10	*bs*	-3.19	-10 (2,37,12)

***Table 4** – Average Value of the preference indicator H_z.*

The Global Wikipedia Network

Based on the value of the preference indicator H_z, we defined a network-like structure over the Wikipedia versions called Global Wikipedia Network – GWN – where version W_a was linked to version W_b if $H_z(a,b) > c_f$, where c_f is a significance level (we used c_f=0.99). This means that W_a is linked to W_b if W_a replicated a significant number of articles from the set $\mathcal{A}_b$ associated with W_b.

The obtained a GWN has a density of 0.159 and average degree of 8.09. The nodes were connected in a single giant weak component and three strong connected components. More precisely, the main component was composed of all the Wikipedia versions except the Dutch and Indonesian versions, which formed two separate components. This implied that the connectedness measure (Krackhardt 1994) was 0.962, meaning that a very large majority of the Wikis could reach each other by a path of any length. Similar to (Travers and Milgram 1969), (Lescovez and Horviz 2008) and (Backstrom 2012), the diameter of the graph was relatively slow (6 paths) compared to the total number of nodes.

Linguistic and Geographical homophily

Our results showed that strong preference mechanisms underpin and bias the hosting relationships between the Wikipedia versions. In this section we explain the presence of such preference mechanisms by arguing that the more similar two countries were (in terms of language and geography) the stronger was the preference between the two correspondent Wikipedia versions. We claim that to explain the emergence of hosting relationships it is necessary to look at the micro social sphere of the motivation that pushed certain authors to translate and share articles in Wikipedia. More precisely, we assumed that the main force driving the Wikipedia contributors to translate and share articles from another specific Wikipedia version was the affinity between their own country and the country of the target Wikipedia version.

We argued that the more similar two countries were, the higher the probability that a relationship of hosting between them was. We referred to the well-known homophily argument (e.g. (McPherson 2001), (Lazarsfeld et al. 1954)) and assumed that the higher the linguistic and cultural similarity between two given Wikipedia language communities , the higher the probability of detecting similarities between them. On the one hand, we modelled the linguistic similarity by defining nine groups of similar languages shown in Table 5. On the other hand, we looked at geographical proximity and defined seven macro areas of similar countries, such as Northern Europe, Southern Europe, Scandinavia, Eastern Europe, Asia, Middle East. To test the hypothesis, we used the well-known E-I index (Lazarsfeld et al. 1954) as a measure of alter-ego similarity across the categories of language and culture. For both language and geographical proximity, we calculated the E-I index of the whole graph, the E-I index of each of the groups and of every single country.

Results showed that hosting processes were driven by the tendency to establish relationships with similar Wikipedia versions, both in terms of language and geographical proximity. Figures 2 and 3 show the network of preferences highlighting geographical and linguistic groups.

ID	Group	Components
1	Ugro-Finnish	*hu, fi, et*
2	Romance	*la, it, es, fr, pt, oc, ca, gl, ro, an*
3	Germanic	*en, de, af, da, is, nl, no, sv*
4	Slavic	*ru, sl, sk, cs, pl, bs, hr, lt, lv, be, bg, sr, uk*
5	Turkish	*az, tr,*
6	Indo-Persian	*fa, ur, hi*
7	Afro-Asiatic	*ar, he*
8	Malay-Indo	*id, ml, ms*
9	Isolated	*sq, th, ml, ka, ko, ja, eu, el, hy, vi, ja, zh*

***Table 5 -** Language Similarity Groups.*

Language. The E-Index of the whole Wikipedia network was 0.052. This indicates a substantial equilibrium between closure to intra-language relationships and openness to extra-language relationships. However, the observed E-I index was significantly lower than the value expected by chance (calculated with a permutation test with 5000 iterations). In other words, while the absolute value of the index indicates a general balance between in-group and extra-group links, the permutation test shows that such intra group relationships were higher than what was expected by chance. This testifies to a tendency towards homophily. Considering the group level E-I index (Table 6), more homophily was present in number 2 (Romance languages) and 4 (Slavic languages), while the others were more open to external relationships.

Group	Internal	External	Total	E-I
1	2	24	26	0.846
2	48	82	130	0.262
3	38	85	123	0.382
4	46	75	121	0.24
5	38	97	135	0.437
6	2	19	21	0.81
7	6	36	42	0.714
8	2	43	45	0.911
9	0	17	17	1

***Table 6 –** The individual E-I index for the inter-intra language relationships. Internal is the number of internal relationships within each of the group; external is the number of external relationships from a group to other groups.*

Lastly, considering the individual E-I index, more homophily was present in *oc* (Occitan) with all the relationships being internal to the group. On the other hand, *hu, uk*, *ml*, and *id* showed complete openness to extra-group relationships, with 0 ties to the Wikipedia versions of their own group.

Group	Internal	External	Total	E-I
1	28	103	131	0.573
2	80	114	194	0.175
3	2	25	27	0.852
4	40	77	117	0.316
5	40	61	101	0.208
6	14	70	84	0.667
7	0	6	6	1

***Table 7 –** The individual E-I index for the inter-intra geographical relationships.*

Geographical proximity. In this case, the E-Index of the whole Wikipedia network was 0.121. This indicates a modest tendency towards openness to extra-group relationships. However, as in the previous case, the observed E-I index was significantly lower than the value that could be expected by chance. The permutation test showed that such intra-group relationships were more frequent than what was expected by chance, thus indicating a tendency towards homophily. Considering the group level E-I index (Table 7), more homophily was present in numbers 2 (Southern Europe), 4 (Eastern Europe) and 5 (Asia), while the others were more open to external relationships.

Wikipedia vs Literature Experts

In the previous section we have shown how articles are not replicated uniformly across Wikipedia versions, but rather are included in each version according to strong preference mechanisms. As a consequence, editors in each Wikipedia community implicitly assign to each set of authors $\mathcal{A}_\ell$ (and the corresponding literature) a different level of importance measured by the number of articles in $\mathcal{A}_\ell$ that are present in their Wikipedia version. For instance, Table 8 shows the top 10 set of authors according to the Finnish Wikipedia W_{fi} and the Turkish Wikipedia W_{tr}. It is evident how each Wikipedia version expresses its own point of view. In the case of these two Wikipedia versions, they both their own authors put at top position, evidence of a local focus in accordance with the findings of the previous section, but they also have 8 out of 10 sets of authors in common, even if their order and relative importance differ. The different size of each Wikipedia version does not affect our analysis, since we are interested in the preference order and the proportion of articles assigned by each Wikipedia community to each set of authors.

Finnish Wikipedia W_{fi}	
Author Set $\mathcal{A}_l$	Articles
fi	2338
en	1222
fr	458
de	333
ru	249
el	201
la	188
es	186
sv	161
it	158

Turkish Wikipedia W_{tr}	
Author Set $\mathcal{A}_l$	Articles
tr	1662
en	565
fr	305
de	245
es	126
it	121
ru	103
el	97
ar	73
la	70

Table 8. *Rank of the set of authors in the Finnish and Turkish Wikipedia versions*

Not only does each of our 52 Wikipedia versions produce a different ranking of the sets of literary authors, but together they also induce a collective ranking, expressing the importance of each set of authors in the entire Wikipedia. More precisely, given a set of authors $\mathcal{A}_\ell$, we can measure its influence $\mathcal{G}(\mathcal{A}_\ell)$ in Wikipedia as a whole by counting the total number of articles about authors in $\mathcal{A}_\ell$ in all the 52 Wiki versions considered. Therefore $\mathcal{G}(\mathcal{A}_\ell)$ is:

$$\mathcal{G}(\mathcal{A}_\ell) = \sum_{a \in \mathcal{A}_\ell} N(a)$$

The resulting ranking is shown in Table 9. For instance, there are 38563 articles about Anglophone authors (set $\mathcal{A}_{en}$) in the 52 Wikipedia considered.

It is no surprise that there are strongly dominant languages both in the individual and in the collective rankings. However, here we are interested in understanding to which degree these rankings could be considered to convene a *fair* point of view or they are mainly a collection of partial opinions about the influence of group of literary authors.

R	$\mathcal{A}_\ell$	$\mathcal{G}(\mathcal{A}_\ell)$	**R**	$\mathcal{A}_\ell$	$\mathcal{G}(\mathcal{A}_\ell)$	**R**	$\mathcal{A}_\ell$	$\mathcal{G}(\mathcal{A}_\ell)$
1	*en*	38563	19	*fa*	2507	36	*is*	979
2	*fr*	24525	20	*ko*	2044	37	*ca*	944
3	*de*	19048	21	*hr*	1877	38	*sq*	926
4	*es*	15937	22	*no*	1800	39	*et*	925
5	*it*	11712	23	*he*	1683	40	*lt*	878
6	*la*	8941	24	*Fi*	1639	41	*bg*	820
7	*ru*	7871	25	*sr*	1529	42	*az*	800
8	*el*	7289	26	*hi*	1507	43	*hy*	689
9	*da*	5620	27	*tr*	1505	44	*ka*	684
10	*ja*	4577	28	*uk*	1503	45	*gl*	521
11	*hu*	3926	29	*ro*	1488	46	*eu*	424
12	*cs*	3892	30	*sl*	1373	47	*bs*	421
13	*ar*	3289	31	*be*	1303	48	*vi*	378
14	*pl*	3114	32	*sk*	1058	49	*id*	370
15	*pt*	3057	33	*ur*	1041	50	*th*	332
16	*nl*	3045	34	*oc*	1040	51	*af*	263
17	*sv*	2602	35	*lv*	1004	52	*ml*	255

Table 9. *The Global ranking of each set of Authors.*

In the absence of an undisputable ground truth about the relative importance of world literatures, we suggest investigating whether the user-generated Wiki ranks are associated with an expert-based rank. We therefore wonder if the resulting importance that each Wikipedia version assigns to each set of authors (and corresponding literature) is comparable to some external expert-based ranking.

The aim of the experiment is ultimately to contrast two ways of generating knowledge: the open and collaborative user-generated Wiki model versus the opinion of a closed and selected group of domain experts.

In order to do investigate this, we collected a list of winners of literary awards and expert-based rankings and we counted how many times each language is mentioned. Our selection criteria were the following: the literary award should be well-established, international – open to any language – and assigned by a panel of experts. We considered the nationality of the winners of the 101 Nobel Prize in literature to date, the 23 winners of the Neustadt Literary Prize and the winners of the Golden Wreath of Struga Poetry Evenings (50 winners since 1966). We also considered the authors included in the list of the *top 100 most influential books* prepared in 2002 by the Norwegian Book Club. The list was based on the opinion of more than 100 authors from 54 countries, asked to nominate the ten books which have had the most decisive impact on the cultural history of the world, including authors from any epoch. By counting how many times a language is mentioned among the 284 list of awarded authors or books, we obtained an expert-based ranking of 32 languages (Table 10).

In order to answer our question we simply need to study the association between the list generated by each Wikipedia version and the expert-based list. We have available not only the ranking of each set of authors in each Wiki, but also a numerical measure of how important each literature is in that Wiki. Since there are significant gaps between the value of $\mathcal{G}(\mathcal{A}_\ell)$ for some set of authors, we suggest to study the Pearson correlation between lists.

Author Set $\mathcal{A}_l$	$\mathcal{M}$	Author Set $\mathcal{A}_l$	$\mathcal{M}$	Author Set $\mathcal{A}_l$	$\mathcal{M}$
1. En	63	12. ar	4	23. fi	2
2. Fr	34	13. la	3	24. sr	2
3. De	25	14. ja	3	25. tr	2
4. Es	24	15. no	3	26. hr	2
5. Ru	19	16. da	2	27. ro	2
6. It	15	17. hu	2	28. oc	1
7. Sv	11	18. hi	2	29. ko	1
8. El	8	19. is	2	30. sl	1
9. Pl	8	20. fa	2	31. bg	1
10. Pt	7	21. he	2	32. sq	1
11. Zh	6	22. cs	2		

***Table 10**. Number of mentions (column $\mathcal{M}$) for each language in a set of expert-based ranks.*

Each Wikipedia version W_x identifies a sequence $A \to \mathbb{N}$ over the ordered list of set of authors A, that a given set of authors $\mathcal{A}_\ell \in A$ returns the number of articles of authors in $\mathcal{A}_\ell$ present in W_x. The Global Wikipedia rank represents an additional list called W_{all}. The expert-based rank defines another continuous variable called $\mathcal{E}$, where a value of 0 is given to all the author sets with no mention in the award list of Table 9. We study the correlation between $\mathcal{E}$ and the 52 individual Wikipedia ranks, and between $\mathcal{E}$ and the global rank W_{all}. We also tested the significance of the person coefficient r with N=50.

Results of our correlation analysis illustrate a clear picture. First, the individual Wikipedia shows a good correlation with the expert-based rank. 25 out of 52 Wikis are correlated at a 0.99 significance level, 8 at 0.95 level and a further 4 at 0.9, meaning that 37 out of 52 Wikipedia versions generate a rank of world literature comparable to that of the experts. The remaining 15 Wikipedia versions that do not correlate with the experts, do not conflict with them (only one marginally negative correlation value was found). The set includes only three major Wiki versions, namely the Japanese, the Danish and the Hungarian – for which we have identified a strong presence of local articles – and minor Wikipedias distributed across Europe, Asia and the Middle East. Despite a strong preference for its own local authors, the majority of Wikipedia versions still correlate with an expert-based ranking. The English Wikipedia has the highest correlation among the major versions, positive evidence of its quality and global status.

Wiki	r	Wiki	r	Wiki	r
W_{all}	0.896 ***	*pl*	0.462 ***	*ml*	0.259*
he	0.771 ***	*ar*	0.418 ***	*be*	0.231*
en	0.741 ***	*it*	0.406 ***	*tr*	0.221
pt	0.729 ***	*zh*	0.39 ***	*sr*	0.193
eu	0.683 ***	*nl*	0.386 ***	*uk*	0.173
la	0.654 ***	*bs*	0.382 ***	*sl*	0.154
fa	0.648 ***	*is*	0.381 ***	*ja*	0.152
ro	0.647 ***	*bg*	0.372 ***	*hi*	0.136
de	0.629 ***	*gl*	0.351 **	*da*	0.129
ru	0.591 ***	*ka*	0.347**	*hr*	0.109
ca	0.589 ***	*et*	0.345**	*th*	0.107
fr	0.584 ***	*ko*	0.335**	*lt*	0.077
oc	0.567 ***	*fi*	0.327**	*hu*	0.055
id	0.535 ***	*vi*	0.32**	*az*	0.055
no	0.527 ***	*sk*	0.302**	*sq*	0.009
sv	0.492 ***	*hy*	0.288**	*ur*	-0.003
af	0.478 ***	*cs*	0.285*		
es	0.471 ***	*el*	0.269*		

***Table 11.** Pearson's correlation coefficient r between Wikipedia Versions and the Experts*

However, the most striking result is the fact that the collective Wikipedia ranking W_{all} has the highest correlation coefficient, ($r \approx 0.9$). Wikipedia exhibits a strong *Wisdom of the Crowds* (Surowiecki 2005) where the collective ranking of all its versions seems to correct their individual biases and generate a list that not only is almost identical to an expert-based one, but that also outperforms all of its components. The linear regression model $W_{all} = \alpha\mathcal{E} + \beta$ (Table 11) showed how one award mention for a language ℓ increments by 615 the total number of articles about authors in $\mathcal{A}_\ell$ in the 52 Wikipedia versions considered.

Coefficients	Estimate	Std. Error	t-value	p-value
α	615.26	0. 02143	28.707	< 2e-16 ***
β	709.29	0. 25649	2.765	0.00794 **
Multiple R-squared: 0.9428, Adjusted R-squared: 0.9417 F-statistic: 824.1 on 1 and 50 DF, p-value: < 2.2e-16				

***Table 12**. Linear regression $W_{all} = \alpha\mathcal{E} + \beta$*

Conclusions

This paper investigated the diffusion of around 100,000 articles among 52 Wikipedia versions. We studied how Wiki versions replicate articles of authors belonging to a particular linguistic group and we collected several findings about the potential mechanisms governing the replication process and its fairness. Results showed that diffusion of articles follows a power law, governed by strong preferences among versions, with a high number of isolated articles. We found that the English Wiki has a prominent role, but also the major European languages had a considerable influence.We also found an important global consensus on classical authors. Latin and Greek authors are in the top ten most replicated set of authors. As shown in Table 4, 50 out of 51 Wikipedias replicate Latin and Greek authors more than expected, and they have the highest value of H_z.We then identified presence of homophilous groups. Our results show that linguistic groups and geographical proximity are significantly correlated with the index of diffusion among Wikipedia versions. Finally, despite the presence of preference mechanisms, we show how the relative importance that each Wikipedia assigns to the set of authors of each language is significantly correlated with an expert-based ranking. However, we believe our main contribution is to have shown how Wikipedia exhibits a solid *Wisdom of Crowds* effect, with the collective ranking of all the Wikipedia versions showing a correlation with the experts higher than any individual Wikipedia version, with a value for Pearson's' *r* of about 0.9.

References

Eom, Y.-H., & Shepelyansky, D. L. 2013. Highlighting Entanglement of Cultures via Ranking of Multilingual Wikipedia Articles. PloS One, 8(10),

NPOV, The Neutral Point of View of Wikipedia, retrieved at en.wikipedia.org/wiki/Wikipedia: Neutral_point_of_view.

Halavais, A., & Lackaff, D. 2008. Analysis of Topical Coverage of Wikipedia. Journal of Computer-Mediated Communication, 13(2), 429–440.

Hecht, B., & Gergle, D. 2009 Measuring Self-focus Bias in Community-maintained Knowledge Repositories. Proceedings of the 4th International Conference on Communities and Technologies - C&T 2009,

Kolbitsch, J., Maurer, H. 2006. The Transformation of the Web: How Communities Shape the Information we consume. Journal of Universal Computer Science 12(2), 187–213.

Warncke-Wang, M., Uduwage, A., Dong, Z., & Riedl, J. 2012. In Search of the Ur-Wikipedia: Universality, Similarity, and Translation in the Wikipedia Inter-language Link Network. Proccedings of WikiSym 2012, Linz.

Denning, Peter, et al. 2005 "Wikipedia Risks." Communications of the ACM 48.12, pages 152-152.

Krackhardt D. 1994, *Graph Theoretical Dimensions of Informal Organizations*, in K. Carley and M. Prietula (eds), Computational Organizational Theory. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. pp. 89-111.

Travers J., Milgram S. 1969, *An Experimental Study of the Small World Problem*, in Sociometry, vol. 32(4), pp.425-443.

Leskovec J. & Horvitz E. 2008, Planetary-Scale Views on an Instant-Messaging Network; available online at http://arxiv.org/abs/0803.0939

Backstrom L, Boldi P, Rosa M, Ugander J, Vigna S 2012, *Four Degrees of Separation*, arXiv:1111.4570v3.

McPherson J, Smith-Lovin L, Cook J 2001, *Birds of a Feather: Homophily in Social Networks*, in Annual Review of Sociology, v.27, pp. 415–444.

Lazarsfeld P.F. and Merton R.K. 1954, *Friendship as a Social Process: A Substantive and Methodological Analysis*, In Freedom and Control in Modern Society. New York, 2001.

Krackhardt D. and Stern R.N. 1988, *Informal Networks and Organizational Crises: an Experimental Simulation*, in "Social Psychology Quarterly", vol.51(2), pp. 123-140.

The 100 most influential books, list retrieved from theguardian.com/world/2002/may/08/books.booksnews on the 15th February 2015

The Neustadt Literary Price, www.neustadtprize.org

The Struga Poetry awards, www.strugapoetryevenings.com

Surowiecki, J. (2005). The wisdom of crowds. Anchor