Wikipedias: Collaborative web-based encyclopedias as complex networks

V. Zlatić,[1] M. Božičević,[2,3] H. Štefančić,[1] and M. Domazet[3]
[1]*Theoretical Physics Division, Rudjer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia*
[2]*Multimedia Institute, Preradovićeva 18, HR-10000 Zagreb, Croatia*
[3]*Institute for Social Research in Zagreb, Amruševa 8, HR-10000 Zagreb, Croatia*
(Received 22 March 2006; published 24 July 2006)

Wikipedia is a popular web-based encyclopedia edited freely and collaboratively by its users. In this paper we present an analysis of Wikipedias in several languages as complex networks. The hyperlinks pointing from one Wikipedia article to another are treated as directed links while the articles represent the nodes of the network. We show that many network characteristics are common to different language versions of Wikipedia, such as their degree distributions, growth, topology, reciprocity, clustering, assortativity, path lengths, and triad significance profiles. These regularities, found in the ensemble of Wikipedias in different languages and of different sizes, point to the existence of a unique growth process. We also compare Wikipedias to other previously studied networks.

DOI: 10.1103/PhysRevE.74.016115 PACS number(s): 05.65.+b, 89.65.−s, 89.20.Hh, 89.75.−k

I. INTRODUCTION

In the last few years the physics community has paid a lot of attention to the field of complex networks. A considerable amount of research has been done on different real world networks, complex network theory, and mathematical models [1–4]. Many real world systems can be described as complex networks: WWW [5], internet routers [6–8], proteins [9], and scientific collaborations [10], among others. Complex network theory benefitted from the study of such networks both from the motivational aspect as well as from the new problems that arise with every newly analyzed system.

In this paper we will present an analysis of Wikipedias in different languages as complex networks. Wikipedia [11] is a web-based encyclopedia with an unusual editorial policy that anybody can freely edit and crosslink articles as long as one follows a simple set of rules. Although there has been a lot of debate on the quality of Wikipedia articles, recent findings reported in Ref. [12] suggest that the factographic accuracy of the English Wikipedia is not much worse than that of the editorially compiled encyclopedias such as *Encyclopaedia Britannica*.

The important facts for this paper are (i) that authors are encouraged to link out of their articles, and (ii) that each Wikipedia is a product of a cooperative community. The former comes in part from the need for lexicographic links providing context for the topic at hand, and in part from the fact that the official Wikipedia article count, serving as the main criterion for comparing encyclopedia sizes, includes only articles that contain an out-link. A community arises initially from the need to follow the central Wikipedia policy of the neutral point of view (NPOV): if there is a dispute regarding the content of an article, effectively all the opposing views and arguments regarding the topic should be addressed. Although there are many occasional contributors, the bulk of the work is done by a minority: roughly 10% of contributors edit 80% of the articles, and the differing degree of authors' involvement serves as a rough criterion for a meritocracy. Hence there is no central structure that governs the writing of a Wikipedia, but the process is not entirely haphazard.

We view each Wikipedia as a network with nodes corresponding to articles and directed links corresponding to hyperlinks between them. There are over 200 Wikipedias in different languages, with different number of nodes and links, which are continuously growing by the addition of new nodes and creation of new links. The model of Wikipedia growth based on the "preferential attachment" [13] has been recently tested against the empirical data [14]. Although different Wikipedias are developed mostly independently, a number of people have contributed in two or more different languages, and thus participated in creating different Wikipedia networks. A certain number of articles have been simply translated from one language Wikipedia into another. Also, larger Wikipedias set precedents for smaller ones on issues of both structure and governance. There is thus a degree of interdependence between Wikipedias in different languages. However, each language community has its unique characteristics and idiosyncrasies, and it can be assumed that the growth of each Wikipedia is an autonomous process, governed by the "function affects structure" maxim.

Namely, despite being produced by independent communities, all Wikipedias (both in their content and in their structure) aim to reflect the "received knowledge" [15], which in general should be universal and interlinguistic. It is expected that community-specific deviations of structure occur in cases where the content is less universal than, e.g., in natural science, but it is also expected that such deviations plague each Wikipedia at some stage of its development. We thus assume we are looking at real network realizations of different stages of essentially the same process of growth, implemented by different communities. By showing which network characteristics are more general and which more particular to individual Wikipedias and the process of Wikipedia growth, we hope to provide insight into generality and/or particularity of the network growth processes.

II. DATA

The main focus of our study is to compare networks of lexicographic articles between different languages. However,

the Wikipedia dataset is very rich, and it is not easily reducible to a simple network in which each Wiki page is a node, as various kinds of Wiki pages play different roles. In particular, the dataset contains

(i) *articles,* "normal" Wiki pages with lexicographic topics;

(ii) *categories,* Wiki pages that serve to categorize articles;

(iii) *images and multimedia* as pages in their own right;

(iv) *user, help,* and *talk* pages;

(v) *redirects,* quasitopics that simply redirect the user to another page;

(vi) *templates,* standardized insets of Wiki text that may add links and categories to a page they are included in; and

(vii) *broken links,* links to articles that have no text and do not exist in the database, but may be created at some future time.

We studied 30 largest language Wikipedias with the data from January 7, 2005. Especially we focused on 11 largest languages as measured by the number of undirected links. In order of size, as measured by the number of nodes, these are English (en), German (de), Japanese (ja), French (fr), Swedish (sv), Polish (pl), Dutch (nl), Spanish (es), Italian (it), Portuguese (pt), and Chinese (zh). Based on different possible approaches to the study we analyzed six different datasets for each language with varying policies concerning the selection of data. We present our results for the smallest subset we studied for each language, designed to match the knowledge network of actual lexicographic topics most closely. It excludes categories, images, multimedia, user, help, and talk pages, as well as broken links, and replaces redirects and templates with direct links between articles. For a detailed explanation of the dataset selection issues, please see our web page [16]. An interesting measurement of the Wikipedia dataset statistical properties is given in Ref. [17], and a nice visualization of the Wikipedia data can be found in Ref. [18].

III. RESULTS

A. Degree distribution

One of the most common features of complex networks is the broad degree probability distribution. The studied Wikipedia networks share this property with many other complex networks, as clearly shown in Fig. 1. The determination of the adequate fitting functional form is a key issue in the analysis of the broad degree distribution. Many complex networks have been found to exhibit the scale free nature characterized by the power-law distribution of node degrees $P(k) \sim k^{-\gamma}$. To investigate a possible power-law behavior, we investigated 11 largest languages. The calculated power-law exponents γ are presented in Table I. To estimate the exponents we used the maximum likelihood formula and a nonlinear fit for the cumulative degree distribution introduced in Ref. [19]. We did not find any significant size effect on the exponents γ. The average γ for different languages is $\gamma_{in} = 2.15 \pm 0.13$, $\gamma_{out} = 2.57 \pm 0.27$, and $\gamma_{und} = 2.35 \pm 0.17$. Calculated average exponents and their standard errors were obtained with the assumption that different realizations of the Wikipedia will have different exponents in the thermodynamical limit. If their values tended to the same limit, standard errors would be smaller as depicted in Fig. 2. While in-degree distributions in general display the power-law behavior, as an example, see Fig. 3, the power-law nature of the out-degree distribution is much less expressed (for an example where the power law is clear, see Fig. 4). Nevertheless, the fat tailed character of the out-degree distribution is beyond doubt. The estimation for the out-degree exponent was calculated in a distant tail where the estimated exponent was sufficiently stable with respect to the minimal degree of the fitted set k_{min}.

In the estimation of average exponents a sample without Polish language values is also considered, as Polish contains spikes related to the calendar pages of the Polish Wikipedia. The decision of the Polish Wikipedia community to heavily interlink calendar pages using standard templates (e.g., the articles for almost every year starting with 5 CE link to all days and months of the year and all years of that century) had enormous repercussions on the degree distribution of the Polish Wikipedia, as can be clearly observed in Fig. 1. The exponents for Polish also differ significantly from other Wikipedia exponents, as can be seen in Table I.

It is interesting to mention that the observed average exponents agree very well with the WWW exponents for Alta Vista reported in Ref. [3].

Alternative distributions we have tested were stretched exponential, log-normal, and the Tsallis distribution. Power law was a significantly better fit than the other distributions with the exception of the Tsallis distribution. Because of the larger number of parameters one needs to estimate for fitting and the unclear phenomenological origin of the Tsallis distribution we decided to report only the power-law exponents which are commonly understood.

Very recently a paper on the Wikipedia network structure [14], by Capocci *et al.*, has appeared. The authors use the complete Wikipedia history to study the growth and structure of Wikipedia as a complex network. In particular, Capocci *et al.* find that the mechanism based on the preferential attachment is adequate for the description of the Wikipedia growth. The paper also analyzes Wikipedia topology and assortativity. The comparison of our results with the results in Ref. [14] for the node degree probability distribution exponents shows an agreement for the in-degree exponents, but reveals a difference in the out-degree exponents (Capocci *et al.* report γ_{out} between 2 and 2.1 whereas our estimated average is 2.6). A possible origin of this discrepancy could lie in the selected dataset of Wiki pages, or in the power-law exponent estimation techniques. Namely, because the out-degree distribution is often not a clear power law, one can expect different results depending on the choice of the minimal degree k_{min} from which one starts the estimation of the power-law exponent, as well as on the choice of the cutoff degree k_{max} up to which a power law is fitted.

The node degree probability distributions, presented in Fig. 1 and Table I, exhibit a high degree of similarity despite the fact that the corresponding Wikipedias differ in size by more than an order of magnitude. This finding supports the assumption that the Wikipedias in different languages represent realizations of the same process of network growth. A

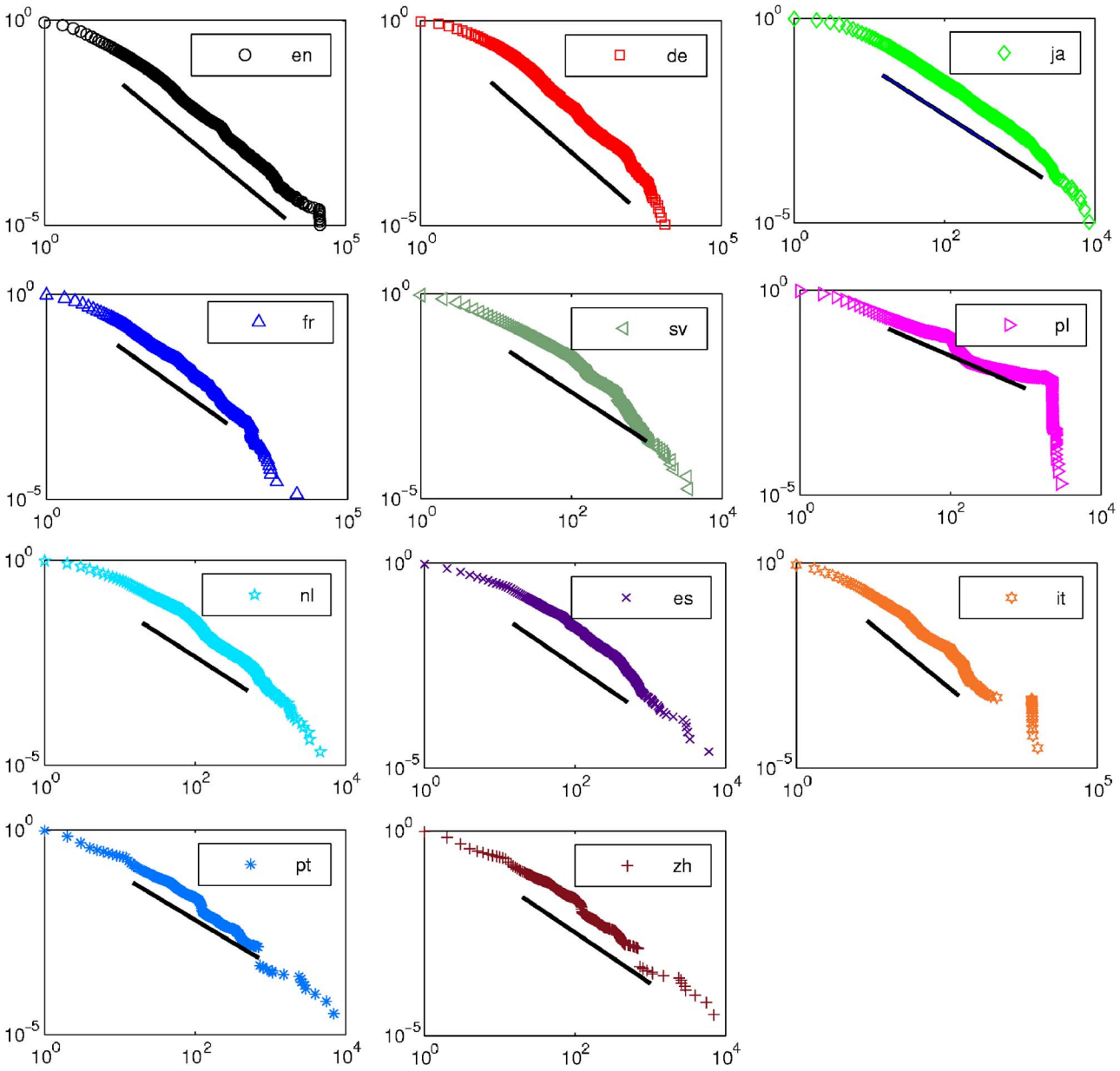

FIG. 1. (Color online) This figure represents cumulative in-degree distributions of the 11 largest languages. In all plots in the figure the abscissa represents the node degree and the ordinate represents the cumulative degree distribution. The start and the end of the drawn best fit straight lines coincide with k_{min} and k_{max} used in fits, respectively. The power law seems applicable to all of them except Polish. This discrepancy is related to the editorial decision of the Polish community to heavily interlink the calendar pages using standard templates. This community decision produced a radical change in the structure of the network. One should also note an unusual distribution for Italian, suggesting a similar cause.

similar claim is expressed by distinguished members of Wikipedian communities [20]. The ensemble of all available Wikipedias thus seems to represent a series of "snapshots" of the Wikipedia growth process. The Wikipedias differ significantly in size and degree of development and therefore the ensemble covers many distinct phases of this growth process.

B. Growth in size

In light of this, we report some interesting features of the growth of the number of crosslinks L with the number of articles N using the said ensemble of Wikipedias. The growth estimated from different Wikipedias is $L \sim N^{\alpha}$ with $\alpha = 1.14 \pm 0.05$, which is close to the linear increase of the number of links with the number of nodes (see Fig. 5). A regular distribution of the points in the plot of Fig. 5 further corroborates the hypothesis of a common growth process. A small difference of the estimated α and 1 is interesting from the perspective of theoretical models aiming to describe complex network growth and structure. Namely, a number of models assume that when a new node is added, approximately the same number of new links are formed. Such models lead to a

TABLE I. The table of γ power-law exponents for in, out, and undirected degree distributions for the 11 largest languages. The exponents for all languages except Polish follow the pattern $\gamma_{out} > \gamma_{undirected} > \gamma_{in}$. It is not a surprise that the Polish language exhibits uncommon behavior having in mind its unusual degree distribution depicted in Fig. 1. The average values and corresponding errors of the universal exponents are calculated in two ways. The upper one is calculated as a mean value and a standard deviation of different exponents in the sample. The lower is calculated with the assumption that all exponents are the same and differences are related to exponent estimation, i.e., the error is calculated as the standard error of the mean. It is important to stress that exponents are not estimated from the degree $k=1$, but from k_{min} for which the estimated exponent is stable.

	In		Out		Undirected	
Language	γ	error	γ	error	γ	error
en	2.21	0.04	2.65	0.15	2.37	0.04
de	2.28	0.05	2.60	0.15	2.45	0.05
ja	2.18	0.03	2.56	0.09	2.41	0.04
fr	2.05	0.06	2.70	0.2	2.38	0.06
sv	2.20	0.1	2.50	0.2	2.30	0.08
pl	1.80	0.1	1.80	0.2	1.85	0.09
nl	2.18	0.12	2.56	0.15	2.38	0.08
es	2.26	0.10	2.70	0.2	2.40	0.08
it	2.20	0.1	2.80	0.2	2.44	0.07
pt	2.10	0.1	2.80	0.2	2.50	0.1
zh	2.24	0.05	2.60	0.1	2.40	0.1
Average	2.15	0.13	2.57	0.27	2.35	0.17
	2.18	0.02	2.58	0.05	2.38	0.02
Average	2.19	0.07	2.65	0.10	2.40	0.05
without pl	2.20	0.02	2.61	0.05	2.40	0.02

linear relationship between L and N and it is interesting that the ensemble of Wikipedias is not far from this linear relationship. Clearly, the models of complex network growth in which the number of links grows with the number of nodes steeper than linearly are also of interest from the perspective of explaining Wikipedia network growth and structure. It would be of special interest to compare the results obtained from the ensemble of Wikipedias with the "snapshots" of a single Wikipedia taken at different stages of its growth. The estimated growth also implies a slight increase of the average degree $\langle k_{dir} \rangle \sim N^{\alpha-1}$. The obtained power-law exponents are greater than 2 and therefore we can expect very limited growth of the average degree, if any.

C. Network topology

In studying the relative sizes of the regions of the network we used a more simplified schema than the taxonomy introduced in Ref. [21] and used in Ref. [14]. We consider two subsets of the network: the giant strongly connected component (SCC), where there is a directed path from every node to another, and the giant weakly connected component (WCC), where there is an undirected path between every two

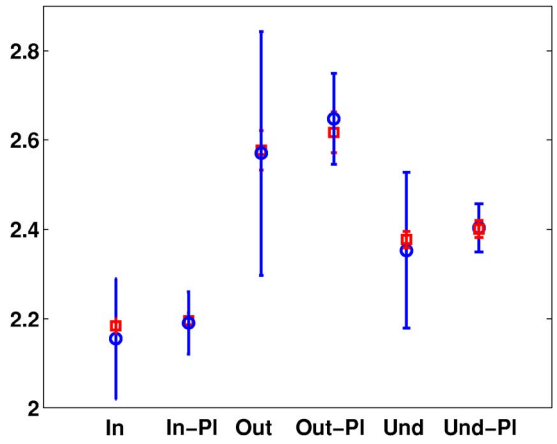

FIG. 2. (Color online) The obtained universal exponents for 11 largest languages. The blue (larger bars) represent mean and the standard deviation of the exponent without the assumption of universality, while the red (smaller bars) represent the standard deviation of the exponents with the assumption of universality.

nodes. The difference between WCC and SCC includes the IN, OUT, TENDRILS, and TUBES components as well as some nodes classified by Ref. [21] as disconnected (DISC). The remaining disconnected nodes are outside the WCC altogether. We present the relative sizes of these regions in Table II. The sizes of the SCC are on the whole larger than ones reported in Ref. [14]. There are two possible ways to account for this difference. First, our dataset could have been built using different criteria of selection. Second, it dates after the introduction of categories to Wikipedia. This was a major structural change, which may have contributed to greater interconnectivity of all lexicographic topics.

D. Reciprocity

Another important characteristic of Wikipedia network is the mutual reciprocity of the links. The average directed degree $\langle k_{dir} \rangle$ is compared with the average undirected degree

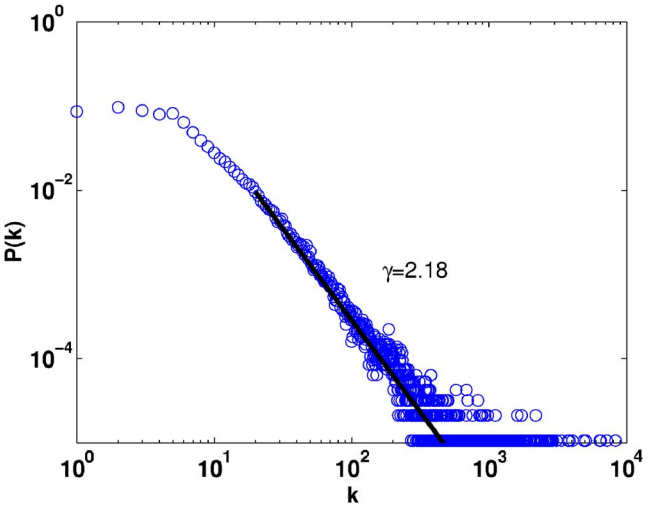

FIG. 3. (Color online) The probability distribution of the in-degree for the Japanese Wikipedia.

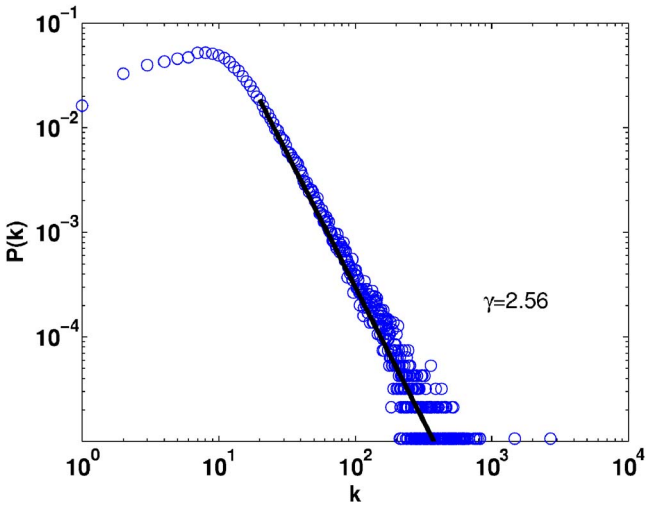

FIG. 4. (Color online) The probability distribution of the out-degree for the Japanese Wikipedia.

$\langle k_{und} \rangle$ in Fig. 6. There is a strong correlation between these two moments. Such correlation leads us to believe that the link reciprocity plays an important role in the Wikipedia growth process. To understand it better we measured unbiased mutual reciprocity using the unbiased measure for reciprocity ρ, presented in the paper by Garlaschelli and Loffredo [22]:

$$\rho = \frac{L_{bd}/L - \bar{a}}{1 - \bar{a}}. \tag{1}$$

Here L_{bd} represents the number of bidirectional links, i.e., links for which a reciprocal link exists. L is the total number of directed links and $\bar{a}$ is the density of the links in the network: $\bar{a} = L/N(N-1)$. The value of reciprocity for the 11 largest Wikipedias is $\rho = 0.32 \pm 0.05$.

It is interesting to compare the reciprocity of Wikipedia with other networks that could be very similar to it. The Wikipedias have a stronger reciprocity than the networks of associations ($\rho = 0.123$ [22]) and dictionary terms ($\rho = 0.194$ [22]), but smaller than the WWW with $\rho = 0.52$ [22]. The difference between the reciprocity of Wikipedia and that of the WWW will be discussed later in the paragraph on the triad significance profile. Small Wikipedias show a decrease in reciprocity with size, which saturates around the reported value, which is very stable for the largest Wikipedias. This stability of the measured value suggests that it is a very important quantity for the description of structure and growth of a Wikipedia-like network.

Reciprocity quantifies mutual "exchange" between the nodes, and can be significant in determining whether and to what degree the network is hierarchical. There have as yet

TABLE II. The table of network components for 11 largest languages, in percentages of the total number of nodes.

Language	SCC	WCC–SCC	All WCC
en	85.73	13.17	1.10
de	95.09	4.63	0.28
ja	96.75	2.77	0.49
fr	94.62	5.01	0.37
sv	89.59	9.36	1.04
pl	93.45	6.00	0.55
nl	94.00	5.69	0.31
es	91.55	7.65	0.81
it	86.12	13.60	0.28
pt	87.73	10.83	1.43
zh	89.01	9.22	1.77

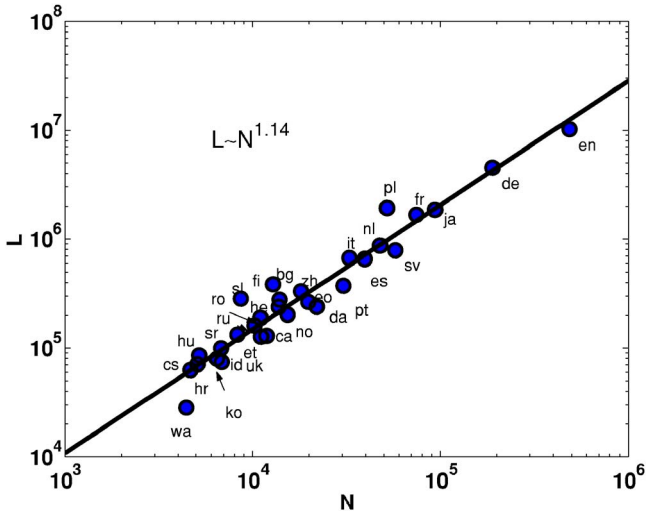

FIG. 5. (Color online) The number of directed links plotted against the number of nodes in different Wikipedias. The growth of L is well described by $N^{1.14}$. This result is very close to a linear relationship and to determine precisely the deviation from linearity, should it exist, the study of the history data for any given language would be necessary.

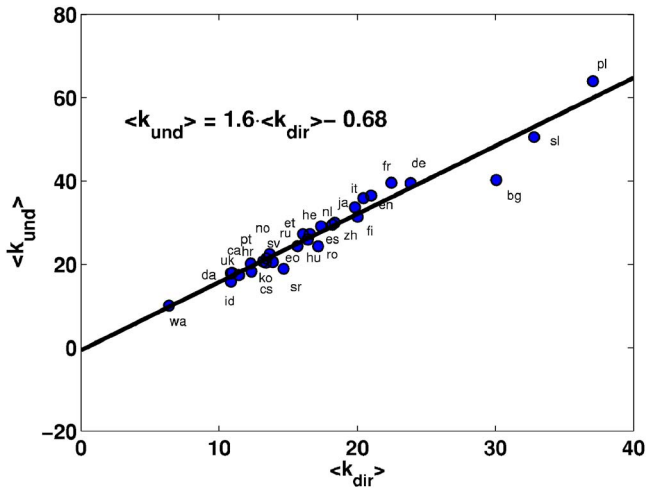

FIG. 6. (Color online) The directed and undirected average degree are in strong correlation across languages. This implies an important and universal characteristic of this measure for the Wikipedia network.

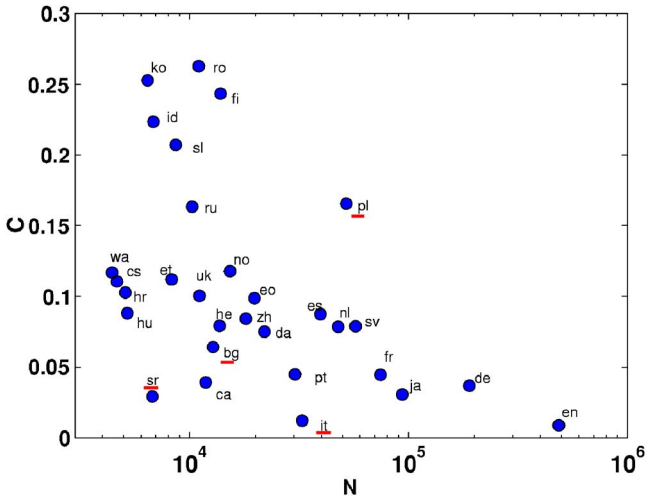

FIG. 7. (Color online) The dependence of the clustering coefficient C on the network size N. Despite the significant scattering of the points, it is possible to argue that Wikipedia clustering coefficient decreases with the network growth. Wikipedias with unusual degree distributions, underlined in red, also exhibit a significant deviation from the trend.

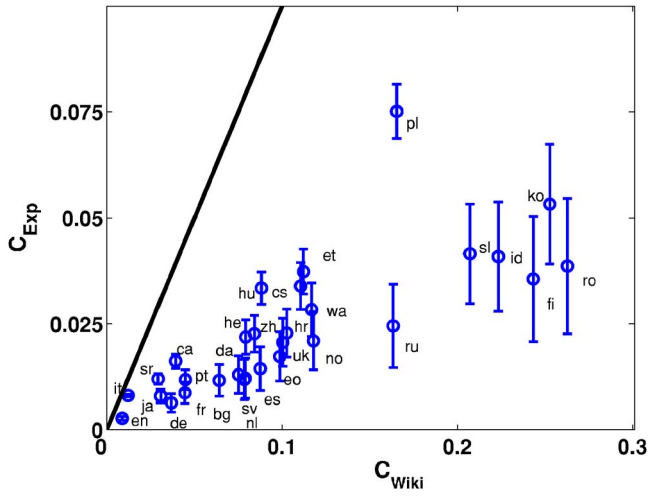

FIG. 8. (Color online) Clustering coefficients of the Wikipedia networks are found to be greater than one would expect from a random network with the same degree distribution. From the figure it is obvious that they cannot be explained as fluctuation from the expected value since the error bars of expected clustering coefficient, calculated as a standard deviation of the sample of randomized networks, are far from the black line which represents $C_{Wiki} = C_{Exp}$. A great diversity of the measured clustering coefficients can be explained by the fact that the original network is directed, and its undirected representation is missing information important for the network growth process.

not been many papers dealing with the origin of reciprocity or network evolution models that capture this quantity.

E. Clustering

The clustering coefficient C is one of the most explored values in complex networks analysis. It is the key quantity in the structure of undirected networks and represents the local correlation effects in the node neighborhood. We calculated the global clustering coefficient, equal to the probability that the two nodes connected with a path of length 2 also have a mutual link, i.e., a path of length 1:

$$C = \frac{3 * \text{number of triangles}}{\text{number of connected node triplets}}. \tag{2}$$

In order to determine the clustering coefficient we regarded the Wikipedia article networks as undirected: every two neighboring nodes are connected with one undirected link. The relation of the clustering coefficients to the network size is displayed in Fig. 7. Although the data points are scattered, the general trend is that the clustering coefficient decreases with the size of the network. This finding is consistent with other results where clustering is a finite-size effect [23]. It is interesting to notice that the points which deviate the most from the general trend, such as Polish or Italian, are also characterized by deformed degree distributions.

We compared the Wikipedia clustering coefficients to the expected clustering coefficients of uncorrelated networks calculated from the known degree probability distribution [23]:

$$C_{exp} = \frac{(\langle k^2 \rangle - \langle k \rangle)^2}{N \langle k \rangle^3}. \tag{3}$$

The peculiarities of Polish, Italian, Bulgarian, and Serbian degree distributions have an enormous impact on this calculation. The expected clustering coefficients obtained by Eq. (3) for Italian, Bulgarian, and Serbian are even greater than 1, which is clearly impossible. These degree distributions exhibit a peak in the ultraconnected nodes, causing a very large second moment $\langle k^2 \rangle$, which spoils the results obtained by analytical reasoning.

An additional contribution to the deviation of Eq. (3) from the empirical values may lie in the fact that the finite maximally random networks with a given degree distribution have some topological constraints (in undirected networks the double links cannot exist, the nodes cannot link to themselves, the sum of degrees has to be even). Therefore these networks are not necessarily uncorrelated and the underlying assumption of Eq. (3) may not be satisfied. It is also plausible that this effect may be more pronounced in networks with slightly pathological distributions.

In order to get a better estimate of the expected clustering coefficient we adapted the algorithm from Ref. [24] for randomizing a network with a known degree distribution, and calculated average clustering coefficients for 100 randomly generated networks. Comparing this clustering coefficient with the measured one, we found a significant bias of the Wikipedia networks to form triangles, see Fig. 8. This is the result one would expect for a network of definitions, because the terms referring to one another are likely to refer to further common terms.

F. Assortativity

We also calculated the assortativity coefficient of the Wikipedia network as a global measure of the degree corre-

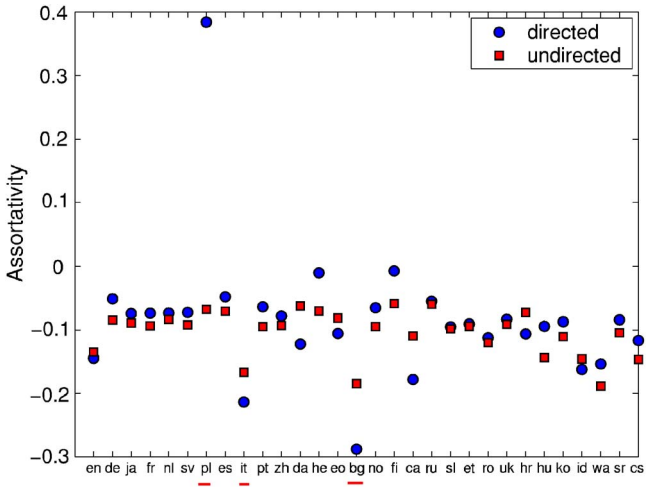

FIG. 9. (Color online) The Wikipedia networks are found to be slightly disassortative on the whole. The outliers are marked with red and coincide with the Wikipedias with peculiar degree distributions.

lations. In Ref. [25] Newman defines the assortativity coefficient r for mixing by vertex degree in a directed network as

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j^{\text{in}} q_k^{\text{out}})}{\sigma_{\text{in}} \sigma_{\text{out}}}. \tag{4}$$

Here e_{jk} represents the probability that a randomly chosen directed link leads out of a node of out-degree k and into a node of in-degree j, q_j^{in}, and q_k^{out} are the degree distributions for in- and outlinks respectively, and σ_{in} and σ_{out} are the standard deviations of these distributions.

This measure describes the likelihood that the nodes of similar (positive values) or dissimilar (negative values) degrees are connected, as compared to the random case. The assortativity coefficient for Wikipedias is slightly negative for all undirected ($r=-0.10\pm0.04$) and directed ($r=-0.10\pm0.05$) Wikipedia networks except the Polish one, which is strongly assortative in the case of the directed network ($r=0.38$), as can bee seen in Fig. 9. The small values of the assortativity coefficient agree well with the more detailed analysis reported by Capocci *et al.* in Ref. [14]. These authors concluded that there was no significant correlation between the in-degrees of the node. Having in mind small values of assortativity coefficient we obtained, this conclusion is very reasonable, but a certain disassortativity is definitely present in Wikipedia because of the overall negativity of almost all measured assortativity coefficients.

G. Path lengths

The path analysis of the Wikipedia networks reveals interesting results, as shown in Table III for the 11 largest languages. The studied quantities are the average path length of the undirected paths in WCC $\langle l_{undir}\rangle$ (calculated as an arithmetic mean) and the average path length of the directed paths in WCC $\langle l_{dir}\rangle$ (calculated as a harmonic mean). For both of these quantities, the largest Wikipedias show no evidence of scaling of the average path lengths with the network size. However, the values of $\langle l_{undir}\rangle$ for all examined networks are close to the expected average path length for a random network $\langle l_{random}\rangle = \ln N / \ln\langle k_{undir}\rangle$, so the Wikipedia networks exhibit small-world behavior in the original sense. In addition, the shortest average path values for the 11 largest languages are very close to one another, with very small scattering around the average value of the sample (see Table III). This scattering is considerably smaller than that of $\langle l_{random}\rangle$.

TABLE III. The table of the average path lengths of the undirected paths in WCC $\langle l_{undir}\rangle$ (arithmetic mean), the average path lengths of the directed paths in WCC $\langle l_{dir}\rangle$ (harmonic mean) and the expected average path lengths for a random network (calculated as $\langle l_{random}\rangle = \ln N / \ln\langle k_{undir}\rangle$), for the 11 largest languages. The displayed average path lengths exhibit no significant dependence on the size of the network despite the fact that the studied Wikipedia networks differ in size by more than an order of magnitude.

Language	$\langle l_{undir}\rangle$	$\langle l_{dir}\rangle$	$\langle l_{random}\rangle$
en	3.28	4.90	3.64
de	3.34	4.33	3.30
ja	3.24	4.10	3.26
fr	3.25	4.36	3.04
sv	3.53	4.84	3.52
pl	3.41	4.47	2.61
nl	3.36	4.40	3.18
es	3.38	4.68	3.20
it	3.11	4.77	2.90
pt	3.35	4.65	3.43
zh	3.26	4.36	2.88
Average	3.32	4.53	3.18
Error	0.11	0.25	0.30

H. Triad significance profile

The last quantity we present in this paper are the triad significance profiles (TSP), introduced in Ref. [24], which describe the local structure of the networks. Counts of specific triads (directed three-node subgraphs, shown in Fig. 10 along the abscissa) in the original network are compared to counts of triads in randomly generated networks with the same degree distribution.

The significance profile SP is the normalized vector

$$SP_i = \frac{Z_i}{(\sum_i Z_i^2)^{1/2}} \tag{5}$$

of statistical significance scores Z_i for each triad i,

$$Z_i = \frac{N_i^{\text{orig}} - \langle N_i^{\text{rand}}\rangle}{\sigma_i^{\text{rand}}}. \tag{6}$$

Here N_i^{orig} is the count of appearances of the triad i in the original network, while $\langle N_i^{\text{rand}}\rangle$ and σ_i^{rand} are the average and

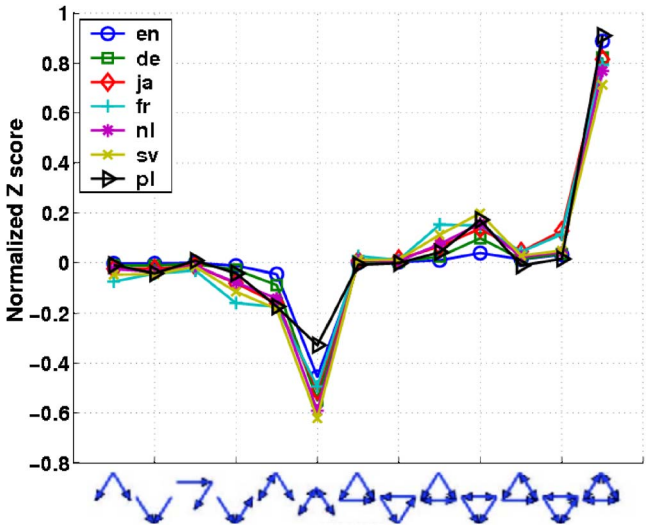

FIG. 10. (Color online) The triad significance profiles of Wikipedias are very similar. The x axis depicts all possible triads of a directed network, while the y axis represents the normalized Z score for a given triad, given by Eq. (5). TSP shapes resemble the TSP of WWW reported in Ref. [24].

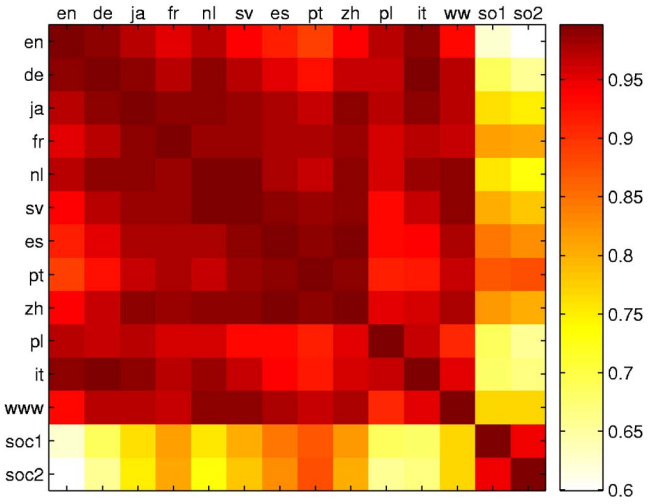

FIG. 11. (Color online) The correlations between TSPs of the 11 largest languages, the WWW of the nd.edu domain [26] and the social networks of positive sentiment between prisoners (soc1) and leadership class students (soc2) [24]. Wikipedias except for Polish and Italian shown in order of size. All Wikipedia profiles and the WWW profile are pairwise very similar. With the exception of Polish and Italian, profiles of languages of similar sizes tend to be more closely correlated. Also, smaller Wikipedias resemble the social networks better than the larger ones do.

the standard deviation of the counts of the triad i over a sample of randomly generated networks.

In Ref. [24], Milo *et al.* identify superfamilies of networks for which triad significance profiles closely resemble each other. Assuming that one can look at the Wikipedia as a representation of the knowledge network created by many contributors, one could expect a possible new superfamily of networks. The triad significance superfamily from Ref. [24] one would expect to be closest to the Wikipedia is the one that includes WWW and social contacts.

The triad significance profile of the largest seven Wikipedias is depicted in Fig. 10, and shows common features found in all examined Wikipedias. These TSPs indeed belong to the same superfamily as the TSPs of WWW and social contacts reported in Ref. [24], see Fig. 11. Within this superfamily, the WWW of nd.edu exhibits higher correlation with the Wikipedias than the social networks do. Since the TSP takes into account the reciprocity of directed links, one could naively expect that Wikipedia reciprocity would also be very similar to the WWW's reciprocity, but we found this is not the case.

The scaling of the triads which are the most represented in the Wikipedia networks (denoted as 10 and 13) with the network size is given in Fig. 12. Since both of these triads represent triangles (see Fig. 10) they contribute to increasing the clustering coefficient. The Wikipedia TSP thus sheds additional light on the large clustering of Wikipedia networks, Fig. 8.

IV. CONCLUSION

We have examined the following characteristics of different language Wikipedia article networks: degree distribution properties, growth, topology, reciprocity, clustering, assortativity, average shortest path lengths, and triad significance profiles. Based on our results, it is very likely that the growth process of Wikipedias is universal. The similarities between Wikipedias in all the measured characteristics suggest that we have observed the same kind of a complex network in different stages of development. We have also found that certain individual Wikipedias, such as Polish or Italian, significantly differ from the other members of the observed set. This difference can be seen most easily in their degree dis-

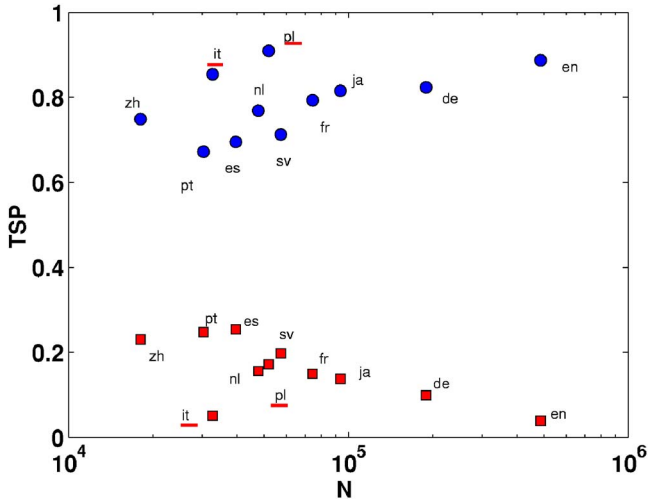

FIG. 12. (Color online) The scaling of the normalized Z score for the most represented triads with the size of the network. The plot demonstrates that the representation of the triad 13 (circles) grows, whereas the representation of the triad 10 (squares) falls with the growth of the network. This effectively means that Wikipedia has a tendency of creating strong (bidirectional) links for the well connected cliques.

tributions, but also shows in assortativity, clustering and the triad significance profile. In the case of the Polish Wikipedia, where the discrepancies are the greatest, we have found that they were caused by an editorial decision involving calendar pages. This shows that the common growth process we have observed is very sensitive to community-driven decisions.

We have shown further that Wikipedia article networks on the whole resemble the WWW networks. Specifically, they belong to the TSP superfamily described in Ref. [24] that includes WWW and social networks, and exhibit small-world behavior, with average shortest path lengths close to those of a random network. In some characteristics, however, large Wikipedias seem to diverge from the WWW. Their reciprocity is lower than that of the WWW reported in Ref. [22], and their average shortest path lengths seem to tend to a stable value.

It is possible that the specific properties of Wikipedias are related to the underlying structure of knowledge, but also that their shared features stem from growth dynamics driven by free contributions, common policies, and community decision making. Whichever the case, the regularities we have found point to the existence of a unique growth process. These findings in turn support the method of using statistical ensembles in network research, and, finally, affirm the role of statistical physics in modeling complex social interaction systems such as Wikipedia.

ACKNOWLEDGMENTS

The work of V. Zlatić and H. Štefančić was supported by the Ministry of Science, Education and Sport of the Republic of Croatia. V. Zlatić would like to thank M. Martinis and the members of his project for the support during the last four years. The authors would like to thank D. Vinković and P. Lazić for important help in computation. We would like to thank R. Milo, N. Kashtan, and U. Alon for making the data and the algorithms from Ref. [24] available on the Weizmann Institute of Science web site. We also thank G. Caldarelli, L. Adamic, P. Stubbs, F. Miličević, and E. Heder for valuable suggestions and discussions, and K. Börner and the Information Visualization Laboratory, Indiana University, Bloomington, for support and cooperation. The work of M. Božičević was partly supported by National Science Foundation Grant No. IIS-0238261.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[2] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[3] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[4] S. H. Strogatz, Nature (London) **410**, 268 (2001).
[5] B. A. Huberman and L. Adamic, Nature (London) **401**, 131 (1999).
[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[7] G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000).
[8] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[9] H. Jeong, S. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature (London) **411**, 41 (2001).
[10] J. W. Grossman and P. D. F. Ion, Congr. Numer. **108**, 129 (1995).
[11] http://www.wikipedia.org
[12] J. Giles, Nature (London) **438**, 900 (2005).
[13] A.-L. Barabási, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).
[14] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, e-print physics/0602026.
[15] W. Roush, Technol. Rev. **Jan 2005**, 21 (2005).
[16] http://www.idd.hr/en/wiki/Wikipedia_project/Dataset_selection.
[17] J. Voss, *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics 2005*, Stockholm, pp. 221–231.
[18] T. Holloway, M. Božičević, and K. Börner, e-print cs.IR/0512085.
[19] M. Newman, Contemp. Phys. **46**, 323 (2005).
[20] J. Wales and M. Rančić (private communication with M. Božičević).
[21] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, S. Stata, A. Tomkins, and J. Weiner, Comput. Netw. **33**, 309 (2000).
[22] D. Garlaschelli and M. I. Loffredo, Phys. Rev. Lett. **93**, 268701 (2004).
[23] S. N. Dorogovtsev, Phys. Rev. E **69**, 027104 (2004).
[24] R. Milo, S. Itzkovtz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Allon, Science **303**, 1538 (2004).
[25] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).
[26] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).