Model of Wikipedia growth based on information exchange via reciprocal arcs

V. Zlatić\textsuperscript{1,2(a)} and H. Štefančić\textsuperscript{1}

\textsuperscript{1} Theoretical Physics Division, Rudjer Bošković Institute - P.O. Box 180, HR-10002 Zagreb, Croatia
\textsuperscript{2} INFM-CNR Centro SMC Dipartimento di Fisica, Sapienza Università di Roma - Piazzale Moro 5, 00185 Roma, Italy, EU

received 17 May 2010; accepted in final form 10 February 2011
published online 11 March 2011

PACS 89.20.Hh – World Wide Web, Internet
PACS 89.65.-s – Social and economic systems
PACS 05.65.+b – Self-organized systems

Abstract – We show how reciprocal arcs significantly influence the structural organization of Wikipedias, online encyclopedias. It is shown that random addition of reciprocal arcs in the static network cannot explain the observed reciprocity of Wikipedias. A model of Wikipedia growth based on preferential attachment and on information exchange via reciprocal arcs is presented. An excellent agreement between in-degree distributions of our model and real Wikipedia networks is achieved without fitting the distributions, but by merely extracting a small number of model parameters from the measurement of real networks.

Copyright © EPLA, 2011

Introduction. – Since lately Wikipedias have been a vibrant interdisciplinary field of study [1–7]. Wikipedia is an “on-line” encyclopedia which is freely edited and written by its contributors. In many cases Wikipedias are represented as networks in which vertices represent Wikipedia pages and directed links (arcs) represent hyperlinks among Wikipedia pages. Interest for Wikipedia and Wikipedia networks in the physicist community is related to the organization of complex networks and general interest for emergence in social systems. Indeed, in the Wikipedia community, any user who feels that he or she can improve the quality of the presented material is free to edit its content, write new articles, create hyperlinks between articles etc. There is no strict central regulatory body that governs the Wikipedia growth and it can be freely said that Wikipedia is a collaborative effort of many independent distinct agents [1]. The unique character of the free-editing article policy and a large number of people participating in the process, make Wikipedia an excellent model system for investigating some complex system ideas in a realistic environment of a real social structure. A key assumption is that different Wikipedias have grown autonomously one from the other. This assumption still has to be verified, but there is a growing body of evidence that this is indeed true [8]. Assuming that this assumption is true, it is truly fascinating that one can find universal network properties, shared by Wikipedia networks written in different languages by different users [2,3]. Such findings are pointing towards the idea that Wikipedias are built through unique processes, which do not depend on the people participating in them and that could be investigated through usual methods of statistical physics. Indeed, in the last few years, there has appeared a growing amount of evidence supporting the use of ideas from statistical physics, graph theory etc. in the description of the social and economic phenomena. This especially applies to phenomena which previously seemed untouchable from the natural scientists point of view [9–11].

One of the very interesting features previously observed in Wikipedias is their reciprocity [2]. Reciprocal arcs are just the arcs pointing from the vertex $i$ to the vertex $j$ for which there is an arc pointing from vertex $j$ to vertex $i$. The reciprocity is then defined as the fraction of reciprocal arcs in the total number of arcs $r = \frac{L^\leftrightarrow}{L}$ [12–14]. It was previously shown that reciprocal arcs can have an interesting role in real networks and in the theory describing them [15–21]. It also seems to be the most stable network measure one can find in the ensemble of Wikipedias except possibly the in-degree distribution exponent [2]. In [22], it was also shown that the reciprocity of Wikipedias cannot be explained by random mixing of arcs. In this paper we show in which manner reciprocal arcs influence the

\textsuperscript{(a)}E-mail: vzlatic@irb.hr
Reciprocity in Wikipedia. – The main aim of this paper is to study how the mechanism of generation of reciprocal arcs affects the structural properties of Wikipedia networks. In some sense, this study is complementary to the research presented in [22] where an opposite approach was adopted, i.e. the effect of various types of correlations on reciprocity was investigated. The simplest assumption on the generation mechanism of reciprocal arcs is that reciprocal arcs are added completely randomly. In particular, for every unidirectional arc in the network one may add a reciprocal arc with some fixed probability p. This model of generating reciprocal arcs and its properties was discussed in detail in [23]. A similar class of models is also investigated in [24]. The most practical way to simulate this model is to take some initial network as a static one and perform the procedure of stochastic addition of reciprocal arcs for each existing unidirectional arc. The most important feature of this model for the purpose of this paper, apart from its simplicity, is that in the model the process of network growth is independent from the process of reciprocal arc formation \(^1\).

An appealing feature of this model of random addition of reciprocal arcs is that it has an analytically tractable effect on degree correlations [23]. Fundamental variables in the model are degrees \(k \equiv (k_i, k_o, k_r)\), where \(k_i\) is the in-degree of the vertex i.e. the number of strictly ingoing edges, \(k_o\) is the out-degree of the vertex i.e. the number of strictly outgoing edges and \(k_r\) is the reciprocal degree of the vertex i.e. the number of reciprocal edges. Particularly, one can use the equation

\[
\langle S(0) \rangle = T^{-1}_{1v}(p)S'(p),
\]

(1)

to study such a process. In eq. (1) \(S'(p)\) represents the vector of product moments of degrees observed in the real network, \(p = \frac{L}{\langle L \rangle}\) represents the fraction of unidirectional arcs that were transformed to bidirectional, \(T^{-1}_{1v}\) is the inverse of the transformation matrix calculated specifically for this process and \(\langle S(0) \rangle\) represents the expected vector of product moments of degrees in the network without reciprocal arcs. These vectors are defined as \(S \equiv \{\langle k_i \rangle = \langle k_o \rangle, \langle k_i^2 \rangle, \langle k_o^2 \rangle, \langle k_r^2 \rangle, \langle k_i k_o \rangle, \langle k_i k_r \rangle, \langle k_o k_r \rangle\}\), while the product moments are defined as: \(\langle k_i k_o \rangle = \sum_{k_i, k_o} P(k_i, k_o)\), where \(P(k_i, k_o)\) is a joint probability distribution that the randomly chosen vertex will have in-degree \(k_i\) and out-degree \(k_o\).

The details of the matrix \(T_{1v}\) are given in [23], but let us just briefly describe it. In the studied model of random addition every unidirectional edge is transformed into a bidirectional one with the probability \(p\). The equation \(P'(k') = \sum_c T(k'|k)P(k)\),

(2)

which expresses a new joint probability that a vertex will have degrees \(k'\) after the addition process is and in the equations above \(T(k'|k)\) is a stochastic matrix of the process which gives the probability that the degrees \(k\) will have degrees \(k'\) after the addition process the degrees of the vertex will be \(k'\). To study changes in product moments one has to calculate for example \(\langle k'|k' \rangle = \sum_{k'_i, k'_o, k'_r} k'_i k'_o P(k')\) and use eq. (2) to connect the resulting moments with initial moments. It turns out that the moment dependences can be represented via matrix \(T_{1v}\).

In fig. 1 we show the correlations in the initial network required to produce the correlations observed in Wikipedia networks for different values of the parameter \(p\). One can see that some types of correlations indicate that the assumption of the growth-independent reciprocal arcs cannot be justified in Wikipedia networks. It is obvious that the parameter \(p\) could not be larger than \(\sim 0.07\) because it would yield negative product moments. On the other hand, the parameter \(p\), extracted from the data, should be around \(\sim 0.16\). This mechanism of reciprocal arcs generation cannot explain the observed reciprocity in Wikipedia networks by itself.

This analysis is based on a very strong assumption that reciprocal arcs are not related to the growth of Wikipedias in any way. On the other hand, it is known that Wikipedia networks grow as many different Wikipedians add and edit many different articles. Clearly it is necessary to investigate the influence of reciprocal arcs on the growth of Wikipedia and vice versa. In [2] we showed that different Wikipedias grow in a very similar fashion and that the

\(^1\)Assuming that the number of reciprocal arcs in the initial network is negligible.
number of newly added arcs is not linear with respect to the number of vertices. Nevertheless, the observed behavior $L \sim N^{1.14}$ is close enough to linear that we can approximate it by a linear growth.

**Model.** – The model used to describe the growth of Wikipedias is studied in detail in [23] and here is presented only the idea of the model and list of the most important analytical results. The studied model was inspired by findings of stability of the reciprocity and degree exponents in different Wikipedia networks [2]. The growth of Wikipedia networks with focus on preferential attachment was previously studied in [3]. In this work a linear-like relationship between the in-degree of the vertex and its probability to acquire a new arc was found, at least for the small and medium vertex degrees, but this study also revealed that there is a significant portion of new arcs forming between vertices which already exist in the network. This very often happens between vertices which are “young” compared to the age of the network [25]. This leads us to believe that ignoring additional formation of arcs between the existing vertices is a reasonable approximation for the growth of the Wikipedia-like network.

The model consists of two steps. In the first one a new vertex, introduced in the network at time $t$, attaches to the network with $m$ outgoing arcs. The probability that the given arc, from these $m$ arcs, will attach itself to some vertex $s < t$ is proportional to the in-degree of the vertex $s$ of the vertex $s$. In the second step for every new arc a new reciprocal arc is formed between vertices $s$ and $t$ with the probability $r$. For such a model it is possible to find exact joint degrees probability distribution $P(k_i, k_o)$ of a single vertex using master equation approach [23]. For example, in the case $m = 1$, the distribution has the form

$$P(k_i, k_o) = \Theta(k_i - k_o)B(k_i, k_o)g(r, k_i, k_o),$$

where $\Theta(x)$ is a usual Heaveside function, $B(k_i, k_o) = \binom{k_o - 1}{k_i}$ and

$$g(r, k_i, k_o) = r^{k_o - 1}(1 - r)^{k_i} \frac{r(1 + r)}{2 + r} \frac{(k_i - 1)!}{(r + 3)_{k_i - 1}}.$$ (4)

In eq. (4) $(r + 3)_{k_i - 1}$ represents Pochhammer symbol [26]. The asymptotic behavior of the marginal in-degree distribution for the described model is of the form

$$P(k_i) \sim k_i^{-(2 + r)}.$$ (5)

This solution nicely interpolates between directed and undirected cases of the BA model [27,28]. Furthermore, the asymptotic behavior of the in-degree distribution given in (5) is also valid for any $m$, i.e. the power-law exponent does not depend on $m$. In [2] were reported the exponent of the in-degree distribution around $\gamma \approx 2.18$ and values of reciprocity coefficient around $r_e \approx 0.35$. The described model for $m = 1$ predicts the relations $r_e = 2r/(1 + r)$ and $\gamma = 2 + r$, which explain the observed empirical values of $r_e$ and $\gamma$.

The three parameters which define the model are: $t$ — the size of the modeled network, $m$ — the number of outgoing arcs of the new vertex and $r$ — the probability of accompanying new arcs with their reciprocal arcs. In order to validate the model, three measured parameters of Wikipedia networks which uniquely describe the degree distributions obtained in the model, were fixed. First the number of vertices in the monitored Wikipedia must be the same as the final size of the model network $i.e. t_{\text{model}} = N_{\text{Wikipedia}}$. In this way it is possible to check if the model also captures the details of the distribution in the tail as well as the power law exponent. The second parameter is the number of arcs in the modeled Wikipedia as the expected number of arcs obtained in the ensemble of model realizations has to be the same as the number of arcs measured in the modeled Wikipedia $i.e. E(\mathbf{L}_{\text{model}}) = L_{\text{Wikipedia}}$. The third parameter is the number of reciprocal arcs as the expected number of reciprocal arcs in the model has to be equal to the measured number of reciprocal arcs $i.e. L_{\text{Wikipedia}}^{\leftrightarrow} = E(L_{\leftrightarrow})$. The last two empirical parameters can be calculated from model parameters in the following way. The number of arcs is

$$L_{\text{Wikipedia}} = E(\mathbf{L}_{\text{model}}) = tm(1 + r),$$ (6)

because $t$ is the number of vertices, $m$ is the number of outgoing arcs every vertex has and $rm$ is equal to the number of outgoing arcs created on old vertices through formation of reciprocal arcs. The number of reciprocal arcs is

$$L_{\text{Wikipedia}}^{\leftrightarrow} = E(L_{\leftrightarrow}) = 2tm\gamma.$$ (7)

From these equations it is easy to express model parameters as functions of measured quantities:

$$m = \frac{L_{\text{Wikipedia}}}{N_{\text{Wikipedia}}^2} - \frac{L_{\text{Wikipedia}}^{\leftrightarrow}}{2},$$ (8)

and

$$r = \frac{L_{\text{Wikipedia}}^{\leftrightarrow}}{2L_{\text{Wikipedia}} - L_{\text{Wikipedia}}^{\leftrightarrow}}.$$ (9)

The parameter $m$ obtained from the measured quantities is not necessarily a natural number, which is an assumption of the analytical treatment [23]. In order to overcome this inconvenience random numbers $\mathbf{m}$ drawn from Poisson distribution $E(\mathbf{m}) = m$, were representing the value of $m$ at any given time. Such a distribution has properties almost identical to our model with $m$ as a natural number if a suitable $E(\mathbf{m})$ is chosen [23].

**Results.** – In fig. 2 an excellent agreement between the in-degree distribution of Japanese Wikipedia and the model distribution is depicted. It is clear that the maximum of the distribution is also well described with the model, which is nice since many of the other models focus only on the exponent of the tail of the distribution.
Fig. 2: (Colour on-line) Comparison of the in-degree distribution of the Japanese Wikipedia with one realization of the model for calculated parameters from eqs. (8) and (9). It is easy to see excellent agreement both between maximum of distribution, and exponent (slope in the log-log plot). The parameters are $t = 94094$, $m = 16.75$, $r = 0.18$. Our simulations show very similar behavior for other studied Wikipedias.

Fig. 3: (Colour on-line) Comparison between cumulative in-degree distribution of English Wikipedia (blue circles) with 100 realizations of the model (red line). The parameters are $t = 486291$, $m = 18.24$, $r = 0.15$. The distribution of the model follows closely the empirical distribution except in the very end of the tail, where the aging effects in the real network could be expected.

It was already mentioned that in [3] small and medium degrees show a kind of preferential attachment. For this reason it is important that the model well describes the maximum which is formed by vertices of a relatively small in-degree. The tail of the distribution is also very well described by the model. It is of paramount importance because such a tail was found to be a universal feature of Wikipedias in different languages.

If one compares a cumulative in-degree distribution of the model to the one of Wikipedias (see fig. 3), one can convince him/herself that the model shows a very good agreement in the tail of the distribution. One can also notice that our model shows a deviation from the monitored Wikipedia at the very end of the distribution. It is not possible to attribute this difference to finite-size effects as the simulated model distribution and data are of the same size. It was already confirmed that the linearity in the attachment principle was observed only for small and medium degrees [3]. The differences in the tail could in principle be explained by aging effects. The largest degrees rarely attract new arcs, because their neighborhood is already matured in its content. We did not model such a behavior in order to keep the model as simple as possible.

The model does not reproduce out-degree distribution well (see fig. 4). The maximum of the modeled distribution is shifted to much to the right and it is too narrow in comparison to the realistic out-degree distributions. This could be the consequence of using Poisson distribution for parameter $m$, which is too narrow in this case. The reason we chose it is just because it is the most easily justifiable one parameter distribution for that case. Clearly, we could get much better results with broader modal distributions for the parameter $m$. In principal, one could expect much broader distribution of the parameter $m$ because different categories of pages in Wikipedia have very different sizes. Since the aim of this paper is to clarify the fundamental role of the reciprocal arcs in the structure and growth of the Wikipedia network, we have focused on the version of the model which requires no additional fitting procedures.

Conclusion. – An excellent agreement of the Wikipedia and model in-degree distribution confirm that the studied model is a natural continuation of the process of preferential attachment, at least for the process of Wikipedia growth. This paper demonstrates a significant value of presented model for understanding of Wikipedia networks and we believe that it could also be important in the case of other types of knowledge networks with time-dependent formation of arcs.

It can be asserted that the presented logic of Wikipedia growth can be attributed to the Wikipedians who are
editing both old and new articles in a very small time frame. In such a case the reciprocity is also a good measure of the information interrelatedness in the knowledge networks. Clearly, the existence of the reciprocal arcs points to a certain intersection of the sets of information presented in different articles. Since reciprocity represents only the first viable correlation for such information sharing, it can be asserted that even better fits could be expected if the model took care of conservation of similar measures such as triad significance profile [29] or some other local structural motives. Such attempts were reported in [30] with interesting results. Taking into account the neighborhood of articles as a pool of more probable information sharing could also improve fitting qualities of the model. The main problem with such attempts is the increase in the number of parameters which such models would require.

Clearly, in the networks which facilitate sharing information, the feedback described by reciprocal arcs can be crucial. In E-mail networks or some other portions of WWW networks, the feedback process of mutual information exchange should have some role. Since reciprocity is a natural representation of feedback, the presented model and its extensions could also be useful in the study of complex systems in which feedback plays an important role. Nevertheless, at present, the usefulness of the presented model in the case of networks of different origin is not clear and further research is needed. The effort in this direction is a logical continuation of this research.

**

This work was financed by the Ministry of Education, Science and Sports of the Republic of Croatia under the contract No. 098-0352828-2863 and by the INFN, Italy. VZ is thankful for the support of G. Caldarelli and would like to thank A. Gabrieleli for reading the manuscript.

REFERENCES

[1] Voss J., Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics 2005, Stockholm 2005, pp. 221–231 http://eprints.rclis.org/handle/10760/6207.
[2] Zlatić V. et al., Phys. Rev. E, 74 (2006) 016115.
[3] Capocci A. et al., Phys. Rev. E, 74 (2006) 036116.
[4] Muchnik L., Itzhack R., Solomon S. and Louzoun Y., Phys. Rev. E, 76 (2007) 016106.
[5] Wilkinson Dennis M. and Huberman Bernardo A., First Monday, 12, issue No. 4 (2007).
[6] Meilir A., in Corpus Linguistics. An International Handbook, edited by Lüdeling A. and Kytö M. (de Gruyter, Berlin, New York) 2007.
[7] Giles J., Nature, 438 (2005) 900.
[8] Suh B. et al., Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations, in VAST ’07: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology (IEEE Computer Society) 2007, pp. 163–170.
[9] Pallotta G., Barabási A.-L. and Vicsek T., Nature, 446 (2007) 664.
[10] Barabási A.-L., Nature, 207 (2005) 435.
[11] De Masi G., Iori G. and Caldarelli G., Phys. Rev. E, 74 (2006) 066112.
[12] Garlaschelli D. and Loffredo M. I., Phys. Rev. Lett., 93 (2004) 268701.
[13] Garlaschelli D., Ruzzzenenti F. and Basosi R., Symmetry, 2 (2010) 1683.
[14] Ruzzzenenti F., Garlaschelli D. and Basosi R., Symmetry, 2 (2010) 1710.
[15] Garlaschelli D. and Loffredo M. I., Phys. Rev. E, 73 (2006) 015101(R).
[16] Catanzaro M. et al., Phys. Rev. E, 70 (2004) 037101.
[17] Serrano M.-A., Boguñá M., Maguitman A. G., Fortunato S. and Vespignani A., ACM Trans. Web, 1 (2007) paper 10.
[18] Boguñá M. and Serrano M.-A., Phys. Rev. E, 93 (2005) 268701.
[19] Zhou Ch., Zemanova L., Zamora G., Hilgetag C. C. and Kurths J., Phys. Rev. Lett., 97 (2006) 238103.
[20] Schnegg M., Int. J. Mod. Phys. C, 17 (2006) 1067.
[21] Kovánen L., arXiv:1002.0763v1 (2010).
[22] Zamora-López G., Zlatić V., Zhou C., Štefančić H. and Kurths J., Phys. Rev. E, 77 (2008) 016106.
[23] Zlatić Vinko and Štefančić Hrovoje, Phys. Rev. E, 80 (2009) 016117.
[24] Rinaldo A. et al., arXiv:1010.0745v1 (2010).
[25] Capocci A., personal communication.
[26] http://mathworld.wolfram.com/PochhammerSymbol.html.
[27] Bolognà B., Borgs C., Chayes J. and Riordan O., Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Baltimore, 2003 (SIAM) 2003, pp. 132–139.
[28] Barabási A.-L., Albert R. and Jeong H., Physica A, 272 (1999) 173.
[29] Milo R. et al., Science, 303 (2004) 1538.
[30] Giammatteo P., Donato D., Zlatić V. and Caldarelli G., EPL, 91 (2010) 18004.