A PageRank-based preferential attachment model for the evolution of the World Wide Web

epl A Letters Journal Exploring the Frontiers of Physics

July 2010

EPL, **91** (2010) 18004 www.epljournal.org
doi: 10.1209/0295-5075/91/18004

A PageRank-based preferential attachment model for the evolution of the World Wide Web

P. Giammatteo[1], D. Donato[2], V. Zlatić[1,3(a)] and G. Caldarelli[1,4,5]

[1] *CNR-ISC and Department of Physics, University of Rome "Sapienza" - P.le Aldo Moro, 5 00185 Rome, Italy, EU*
[2] *Yahoo! Research Lab - Barcelona, Spain, EU*
[3] *Theoretical Physics Division, Rudjer Bošković Institute - P.O. Box 180, HR-10002 Zagreb, Croatia*
[4] *Linkalab Complex Systems Computational Lab. - 09100 Cagliari, Italy, EU*
[5] *London Institute of Mathematical Sciences - 22 South Audley St Mayfair, London W1K 2NY, UK, EU*

received 7 December 2009; accepted in final form 25 June 2010
published online 28 July 2010

PACS 89.75.Da – Systems obeying scaling laws
PACS 89.75.Hc – Networks and genealogical trees
PACS 89.20.Hh – World Wide Web, Internet

Abstract – We propose a model of network growth aimed at mimicking the evolution of the World Wide Web. To this purpose, we take as a key quantity, in the network evolution, the centrality or importance of a vertex as measured by its PageRank. Using a preferential attachment rule and a rewiring procedure based on this quantity, we can reproduce most of the topological properties of the system.

In recent years we have witnessed an incredible growth in size and complexity of some technological systems as the World Wide Web. Due to its long exponential growth, the size of the WWW has grown to such an extent that, nowadays, it is possible to apply standard methods of Statistical Physics in order to describe it. For this reason the field of Complex Networks has been originated in the last century from the first analysis of the WWW [1,2]. Since then it has developed in order to model the growth of various technological systems [3–6] and to describe the topology and growth of various other systems ranging from Collaborative Systems [7,8], to Biology [9] and Finance [10,11]. The WWW has been attracting the interest of many different scientific communities in the attempt to understand how it evolves. This effort has determined the flourishing of several generative models [12–14] devoted to capture some of the most common properties of the Web [15]. Here we present a contribution aimed at discovering the microscopic (at user level) forces shaping these self-organized structures.

The Barabási-Albert (BA) model was historically the first one to reproduce one of the experimental evidence of these systems, namely the power-law distribution of the node degree [12] (the number of link connected to a specific node). In this model, at every time step, new vertices enter the system and they are connected to the old ones by drawing a fixed number m of vertices to them. These end-vertices are chosen with a probability given by their degree. It was shown that the BA model produces networks where the degree k of a vertex is distributed according to a power law $P(k) \sim k^{-\gamma}$ with an exponent γ equal to 3 (in analogy with the values observed in the various real networks). In the case of directed networks as the WWW, where the edges (in this case the hyperlinks) are directed, *i.e.* can be followed only in one direction, one deals with both in- and out-degree (edges pointing in and edges pointing out).

Real networks are also characterized by various topological properties often related to the specific system under consideration. For example, the WWW is known to have a "Bow Tie Structure" (BTS) [16,17] (see fig. 1) where some vertices are mutually reachable one from another (the Strongly Connected Component SCC), some allow to enter into the SCC (and they form the IN component), others are reachable from the SCC and form the OUT component [12,18].

Experimentally, the WWW is known to have a power-law in-degree exponent $\gamma_{in} = 2.1$, a power-law out-degree exponent $\gamma_{out} = 2.3$ (even if some recent analysis reports the existence of cut-offs [19]), a value of the SCC of about 60%–70% and an IN and OUT components of similar size for about 20%–15% of the total.

(a) E-mail: vzlatic@irb.hr

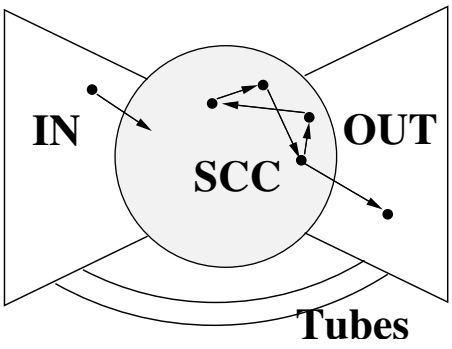

Fig. 1: The bow-tie structure of the WWW. On top of IN, OUT and SCC, there are also pages directly connecting IN and OUT (Tubes); pages pointing nowhere (Tendrils) as well as disconnected components, the latter two kinds of pages are not shown in this picture and account for a very small proportion of the web.

Furthermore, in the WWW it is possible to assess the importance or centrality of a particular vertex by measuring its PageRank [20]. This measure is the core of the success of the search engine Google, founded to rank value of the pages on the web. The PageRank, $PR(i)$, of a vertex i is defined through the equation

$$PR(i) = \alpha \sum_{j \to i} \frac{PR(j)}{k^{out}(j)} + \frac{\alpha}{N} \sum_{j \in D} PR(j) + \frac{1-\alpha}{N} \qquad (1)$$

and can be considered as the fraction of time spent by a random walker on the particular vertex i (D is the set of nodes which do not have out-going links). Since this quantity is commonly considered a good measure of popularity of a web page, we consider it also a good measure of the probability to attract future hyperlinks. For this reason, we define here a model where the preferential attachment mechanism is based on the PageRank quantity.

In-degree and PageRank are known to be correlated (at least in a certain interval of in-degree values) [21], nevertheless the results of the two models differ from each other. The choice of this "microscopic" mechanism of growth coupled to a rewiring rule, reproduces nicely some of the properties of the WWW. Furthermore since PageRank determines the traffic on selected pages by putting them upfront in search engines, this mechanism could represent one of the simplest way to couple topology and dynamics on a network in the spirit of self-organized models [22]. Finally, since PageRank is often interpreted in terms of random walkers on the graphs, this model can be compared with other models [23–25] using other choices of preferential attachment rules.

In the classic Barabási-Albert (BA) model new vertices receive incoming links with a probability proportional to their degree. More recently [14] proposed a model that complements the BA model by choosing the end-point of a link with probability proportional to the in-degree and to the PageRank of a vertex. There are two parameters $\alpha, \beta \in [0,1]$ such that $\alpha + \beta \leqslant 1$. With probability α a node is chosen as the end-point of the the l-th edge with probability proportional to its in-degree (preferential attachment), with probability β it is chosen with probability proportional to its PageRank value, and with probability $1-\alpha-\beta$ at random (uniform probability). The authors show by computer simulation that with an appropriate tuning of the parameters the generated graphs capture the distributional properties of both PageRank and in-degree. We refer to this model as the *PRU* model.

Also in our model the key quantity that determines the probability of old vertex acquiring a new link is instead the PageRank. In other words, the larger is the PageRank value of a vertex, the larger is the probability to receive further links from the newly added vertices. Contrarily to both the original BA and *PRU* model, this model allows the use of directed edges whose proportion is ruled out by a model parameter; finally, in the spirit of generalized BA model [26], also the possibility of rewiring between different existing vertices is taken into account. Since this fundamental modification, the structure arising from this model contains cycles and SCCs. All the previous models do not encompass an explicit mechanism for the rewiring and do not contains SCCs. In order to solve this major drawback an artificial rewiring is arbitrarily introduced after the generation of the graphs adding new links or inverting the direction of a fraction of the links. Our model solves this major drawback considering the rewiring as part of the generative mechanism.

While, *macroscopically*, data analysis shows that, at least in a finite region, the in-degree of a vertex and its PageRank are correlated, it is interesting to consider a model where PageRank appears as a *microscopical* rule of growth.

We start from an initial network of few (*i.e.* 3–5) vertices and edges representing the network at time $t = t_0$. Time is discretized and at every successive time step t, there are two possibilities: either a new node is introduced with probability P_r (node growth) or (with the complementary probability $(1-P_r)$) a new link is created in the network.

In the first case (new vertex), the new node brings a certain number m of new edges. The end-vertices are chosen according to their PageRank value. The parameter m is a random variable chosen from a uniform probability distribution $F(m)$, defined in the interval $[1; m_0]$, where m_0 is the number of nodes present in the initial core (to avoid creating a number of edges more than the available vertices). Once the edge has been created, the direction is chosen out-going from the new vertex with probability P_v and in-going with the complementary probability $(1-P_v)$.

In the second case (new edge), a new link is added between a pair of existing unconnected vertices. Both of them are chosen with a probability related to their PageRank. Actually the two probabilities, P_{ro} for the out-going and P_{ri} for the in-going node, are not exactly those one can obtain from the PageRank distribution. This is due to the necessity of excluding the couples of vertices

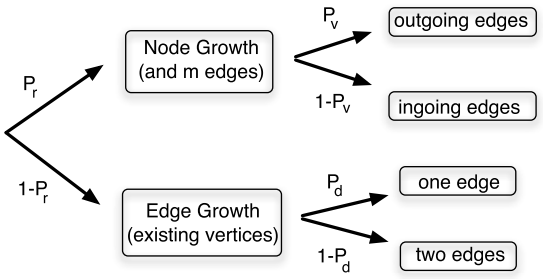

Fig. 2: A sketch of the model growth.

already connected each other. Therefore, if we take the P_{ro} as the value of PageRank for the origin vertex, we have to compute P_{ri} from the subset of vertices not connected with the first one. With probability P_d we also add the opposite edge (if it already does not exist), therefore creating a bidirectional link between the vertices; with probability $1-P_d$ instead we keep only one edge. At the first time step, after introducing a new link or a new node, we calculate the new PageRank value of the nodes using the *power method*. During the iterations, we instead use the *adaptive method* [27] to accelerate the convergence to the PageRank stationary point.

The three parameters describing the evolution of this model (see fig. 2) are therefore P_r, P_v and P_d. As expected, we do not notice any dependence upon the value of the initial nodes m_0. Actually, various m_0 values were taken as initial conditions (from 2 to 7), but neither the decay exponent γ nor the BTS show significant differences in their values. Since also PageRank is defined by assigning a defined "teleportation" probability, our model does depend upon the parameter α as well. In most of the simulations we kept the original value assigned in the original PageRank paper.

In the numerical simulations of the model we made an exploration of the parameter space defined by the three probabilities P_r, P_v, P_d checking:

– the degree distributions of the in and out degree are power-law functions with the same decay exponents observed in other works [12,18];

– the network topology is structured in order to identify the BTS inside the network itself as showed in several papers [16,17].

For every choice of the three parameters, we produced an ensemble of 100 networks each composed by 10000 vertices.

Before proceeding to see what happens for different values of the parameters P_r, P_v and P_d, we consider two limit cases which reproduce two well known networks. The first one is obtained by setting $P_r = 0$ and $P_d = 1$ (P_v can be any when $P_r = 0$), which reproduce a situation similar to the Random Graph model of Erdős-Rènyi (ER) [28] obtaining a Poisson distribution both for in-degree and out-degree.

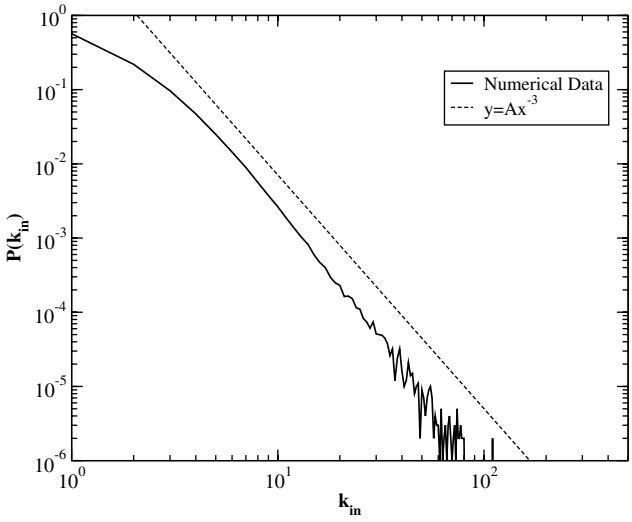

Fig. 3: In-degree distribution in the BA limit with $P_r = 1$ and $P_v = 1$; this is an ensemble of 100 networks of 10000 nodes. Here the decay exponent is $\gamma = 3.01$ and the error, obtained with the likelihood method, is $\sigma = 0.02$.

The second case should instead mimic the Barabási-Albert (BA) Model, whose dynamic is obtained by setting $P_r = 1$ and $P_v = 1$ (P_d here is not important if $P_r = 1$). Note that while the dynamic is similar, the microscopic quantity in the preferential attachment is not, since we use the PageRank instead of the degree. Therefore different choices of α can trigger different quantitative results. Indeed with this choice of the parameters (for which the network can be considered as undirected) and a value of $\alpha \simeq 0.85$, the degree distribution is a power law with a decay exponent $\gamma = 3.01 \pm 0.02$ (see fig. 3 where fit has been done according to the likelihood method [29]), in good agreement with the value expected in the BA Model ($\gamma = 3$). Instead, when $\alpha = 1$ and no randomness is present (in the form of teleportation probability) the results are quantitatively different. That is different α values produce still the power-law degree distribution with different values of the exponent.

Exploring the other possible parameters choices, (provided we are enough far away from the ER limit) we obtain a whole series of power-law distributions differing for their exponents. A sketch of a typical case is shown in fig. 4.

Indeed, it is possible to observe a power-law function both for in-degree and out-degree distribution, some really similar to the real ones of the WWW, with decay exponents close to the experimental ones. For instance, values of $P_r = 0.3$ $P_v = 0.9$ $P_d = 1$ produce decay exponents of $\gamma_{in} \simeq 2.1$ and $\gamma_{out} \simeq 2.7$, closer to the ones observed in the real data [12,18]. In a recent work [19], WWW out-degree distribution seems to be dominated by exponential cut-off, in contrast with our model and previous Web analysis. Anyway we remember that the cut-off effects depends by a number of factor like on how much costly it is the

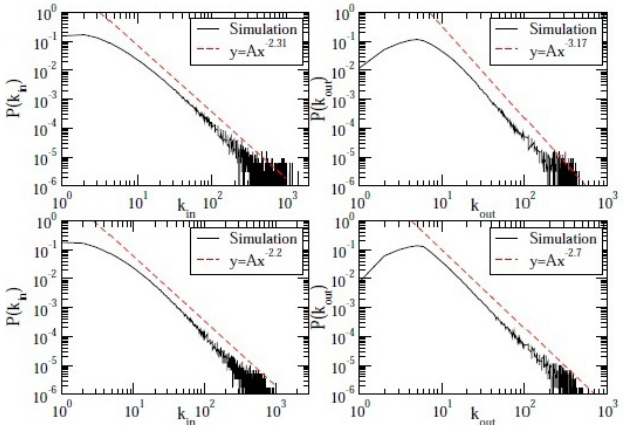

Fig. 4: (Colour on-line) The upper two panels represent in-degree (left) and out-degree (right) distributions for $P_r = 0.2$, $P_v = 0.8$ and $P_d = 0.4$. The lower two panels represent in-degree (left) and out-degree (right) distributions for the values $P_r = 0.3$, $P_v = 0.9$ and $P_d = 1$.

Table 1: Size (percentage on the total) of the various BTS zones with respect to parameters value. Here we show the three main parts of the BTS: SCC, IN and OUT components. With TTD we intend the rest of the BTS (Tubes, Tendrils and Disconnected components).

P_r	P_v	P_d	SCC	IN	OUT	TTD
0.9	0.3	1.0	77.53	6.33	15.88	0.26
0.8	0.8	0.4	56.59	40.37	2.71	0.33
0.8	0.4	0.4	78.66	9.48	11.48	0.38
0.5	0.4	0.8	85.06	6.34	8.47	0.13
0.5	0.4	0.4	83.54	7.18	9.10	0.18
0.2	0.8	0.8	72.46	25.58	1.88	0.08
0.2	0.8	0.4	55.30	41.80	2.60	0.30
0.2	0.4	0.8	92.74	2.81	4.43	0.02

generation of out-degree. Furthermore, for particular kind of networks the cut-off is also observed in the in-degree distribution. According to these researches [30,31] the cut-off effects seems to disappear when the number of edges increase and the network become less sparse. Similarly, a cut-off is not observed in Wikipedia where the average number of out-degrees per length of page is on average higher than that for normal Web.

We then passed to analyze the parameter space with respect to the Bow-Tie Structure [16,17] in the network topology. By considering the results summarized in table 1, we note that the SCC size is mostly governed by the parameter P_r. In particular, the smaller is P_r, the bigger is the size of the SCC zone.

A large value of reciprocity probability P_d accelerates this process. On the other hand, P_v regulates the dimension of the other two zones of the BTS: the IN and the OUT component. Indeed, P_v determines the orientation of the new links introduced by the new node. For example, the smaller its value and the bigger the OUT component.

Unfortunately experimental data [17,18] on the real dimension of the zone in WWW are not univocal. By comparison with the similar (but more compact case) of Wikipedia we expect a large SCC zone (the largest component of the structure with values around the 70%). The remaining is divided mostly between the other two set, IN and OUT. The other zones together, usually take only a small proportion of the network. It is a well-known fact that the proportions of the BTS main components seem to be dependent on the crawl considered. If the seed of the crawl is not properly chosen we can not discover at all the nodes in the IN component. The relative dimensions depends also by the maturity of the network and it is proved that nodes migrate from partition to partition as time evolves [32]. Actually the SCC becomes larger as consequence of the densification. To avoid the problem of the crawl dependence as much as possible, we measured the BTS zone dimensions taking randomly from the graph finally builded, a node and then "burn" the graph forward and backward respect to the chosen vertex. We collected 100 nodes randomly, we burned forward and backward the graph and then we intersected the sets of nodes obtained for each node. The intersection of these sets is the SCC, the backward node set excluded from the intersection is the IN component and the forward node set excluded is the OUT component (the rest account for Tendrils, Tubes and Disconnected). Within these 100 nodes we choose also the one with the highest degree to increase our probability to reach the right values of the BTS zones.

In most of the cases the model reproduces a BTS with a large SCC and two smaller IN and OUT components. More particularly, by tuning the parameter values of P_r, P_v and P_d to reproduce the observed degree distribution exponents, the model produces a SCC core with a dimension larger than 70% and an OUT component smaller than the IN component.

Different choices are possible as shown in fig. 4, *i.e.* the case $P_r = 0.2$, $P_v = 0.4$ and $P_d = 0.8$ which shows a nice agreement with the experimental BTS (see table 1), an out-degree decay exponent $\gamma_{out} \simeq 2.8$ and in-degree decay exponent $\gamma_{in} \simeq 2.4$.

Finally, we considered the structure of bipartite cliques present in the model. As previously shown in [33], these structure accounts as elementary elements for the formation of communities and, as much as the cycles [34], they account for the robustness of the system. They are also particularly important for the validation of WWW models. This kind of structure could help to find communities in a graph [35] and maybe could also help to see if the model proposed is able to reproduce groups of nodes which share the same information and topology. Almost any model, apart from the copying one [25], does not naturally form bipartite cliques (BC) despite their presence in the real WWW. In our model we find an exponential decaying distributions of the number of bipartite cliques (i, j). This quantity has been computed by using a semi-external

algorithm for computing bipartite cores. The algorithm in [36], is an external one that stores the graph in secondary memory in a number of blocks. Every block $b, b = 1, \ldots, [N/B]$, contains the list of successors and the list of predecessors of B vertices of the graph. The idea is that, at each iteration, we load in the main memory a single block, and we try to compute all the bipartite cliques (BC) inside that one; at the same time, we keep track (in a buffer file), of the partially computed BC, *i.e.* the ones that could become BC depending on the part of the graph that either we do not have seen yet or we have seen without knowing it could have been part of a BC. Our algorithm needs to bring in main memory each block at most twice; the first when it builds the partially computed BC, and the second when it looks for the missing parts. By focussing on the clique (3,4) we find that their presence in the graph is exponential distributed and therefore slightly smaller than the real case. We comment that feature by noting that our simplified model does not take into account all the rewiring rules taking place in the WWW which makes it more compact and more reciprocal. So our model does not reproduce successfully a community structure.

In summary, we presented a simplified model of WWW growth, that even if inspired to the traditional BA model, is based on a quantity related to WWW development and growth. Despite its simplicity, the model reproduces the power-law function for both the in-degree and the out-degree distributions. Also, it gives non-trivial values for the BTS components.

REFERENCES

[1] Albert R., Jeong H. and Barabási A.-L., *Nature*, **401** (1999) 130.
[2] Huberman B. A. and Adamic L., *Nature*, **401** (1999) 131.
[3] Faloutsos M., Faloutsos P. and Faloutsos C., *ACM SIGCOMM '99*, **29** (1999) 251.
[4] Pastor-Satorras R. and Vespignani A., *Evolution and Structure of the Internet* (Cambridge University Press) 2004.
[5] Zlatić V., Božičević M., Štefančić H. and Domazet M., *Phys. Rev. E*, **74** (2006) 016115.
[6] Capocci A. *et al.*, *Phys. Rev. E*, **74** (2006) 036116.
[7] Newman M. E. J., *Proc. Natl. Acad. Sci. U.S.A.*, **98** (2001) 404.
[8] Catanzaro M., Caldarelli G. and Pietronero L., *Phys. Rev. E*, **70** (2004) 037101.
[9] Han J.-D. J. *et al.*, *Nature*, **430** (2004) 88.
[10] Bonanno G., Caldarelli G., Lillo F. and Mantegna R. N., *Phys. Rev. E*, **68** (2003) 046130.
[11] Battiston S. *et al.*, *J. Econ. Dyn. Control*, **31** (2007) 2061.
[12] Barabási A.-L. and Albert R., *Rev. Mod. Phys.*, **74** (2002) 47.
[13] Kumar R., Raghavan P., Rajagopalan S., Sivakumar D., Tomkins A. and Upfal E., *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science, Redondo Beach, CA, 12–14 November 2000,* pp. 57–65, ISBN: 0-7695-0850-2.
[14] Pandurangan G., Raghavan P. and Upfal E., *Internet Math.*, **3** (2006) 1.
[15] Costa L., da F., Rodrigues F. A., Travieso G. and Villas Boas V. R., *Adv. Phys.*, **56** (2007) 167.
[16] Broder A. *et al.*, *Comput. Netw.*, **33** (2000) 309.
[17] Donato D., Leonardi S., Millozzi S. and Tsaparas P., *J. Phys. A*, **41** (2008) 224017.
[18] Caldarelli G., *Scale-Free Networks* (Oxford University Press) 2007.
[19] Serrano M., Maguitman A., Boguñá M., Fortunato S. and Vespignani A., *ACM Trans. Web*, **1** (2007) Article No. 10.
[20] Brin S. and Page L., *Comput. Netw.*, **30** (1998) 107.
[21] Fortunato S. and Flammini A., *Int. J. Bifurcat. Chaos*, **17** (2007) 2343.
[22] Garlaschelli D., Capocci A. and Caldarelli G., *Nat. Phys.*, **3** (2007) 813.
[23] Lambiotte R. and Ausloos M., *EPL*, **77** (2007) 58002.
[24] Saichev A. and Sornette D., *Phys. Rev. E*, **72** (2005) 026138.
[25] Kleinberg J., Ravi Kumar S., Raghavan P., Rajagopalan S. and Tomkins A., *The Web as a Graph: Measurements, Models and Methods*, in *Computing and Combinatorics: Proceedings of the 5th Annual International Conference (COCOON '99)*, *Lect. Notes Comput. Sci.*, Vol. **1627** (Springer, New York) 1999, pp. 1–17.
[26] Krapivski P. L., Rodgers G. J. and Redner S., *Phys. Rev. Lett.*, **86** (2001) 5401.
[27] Berkhin P., *Internet Math.*, **2** (2005) 73.
[28] Erdős P. and Rényi A., *Publ. Math. (Debrecen)*, **6** (1959) 290.
[29] Newman M. E. J., arXiv:cond-mat/0412004v3 (2006).
[30] Amaral L. A. N., Scala A., Barthlmy M. and Stanley H. E., *Proc. Natl. Acad. Sci. U.S.A*, **97** (2000) 11149.
[31] Leskovec J., Kleinberg J. and Faloutsos C., in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, *Chicago*, *Illinois*, *USA* (ACM Press, New York) 2005, pp. 177–187.
[32] Buriol L. S., Castillo C., Donato D., Leonardi S. and Millozzi S., *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (ACM Web Intelligence) 2006, pp. 45–51.
[33] Donato D., Laura L., Leonardi S. and Millozzi S., *ACM Trans. Internet Technol. (TOIT)*, **7** (2007) Article No. 4.
[34] Bianconi G., Caldarelli G. and Capocci A., *Phys. Rev. E*, **71** (2005) 066116.
[35] Fortunato S., *Phys. Rep.*, **486** (2010) 75, DOI: 10.1016/j.physrep.2009.11.002.
[36] Donato D., Laura L., Leonardi S. and Millozzi S., A software library for generating and measuring massive webgraphs. Technical Report D13 COSIN European Research Project, `http://www.dis.uniroma1.it/cosin/html_pages/COSIN-Tools.html`.