Properties of a Growing Random Directed Network

G.J.Rodgers* and K. Darby-Dowman

Department of Mathematical Sciences, Brunel University

Uxbridge, Middlesex UB8 3PH, U.K.

**Electronic address: G.J.Rodgers@brunel.ac.uk*

Abstract

We study a number of properties of a simple random growing directed network which can be used to model real directed networks such as the world-wide web and call graphs. We confirm numerically that the distributions of in- and out-degree are consistent with a power law, in agreement with previous analytical results and with empirical measurements from real graphs. We study the distribution and mean of the minimum path length, the high degree nodes, the appearance and size of the giant component and the topology of the nodes outside the giant component. These properties are compared with empirical studies of the world-wide web.

PACS numbers: 02.50.cw, 05.40.-a, 89.75Hc.

I. INTRODUCTION

Understanding the geometry and evolution of complex distributed networks is a major challenge for statistical physics [1,2]. These networks are very important technologically, and many of them display power law distributions of node degree, a feature not found in standard random graphs [3]. There are a number of important questions to answer. These include how to make use of their unusual connectivity to create efficient search and crawl algorithms [4], their resistance to random or intentional attack [5,6], how they will evolve over time, and how we can manage their geometry to make them more efficient [7,8].

The www graph is the directed graph whose nodes correspond to pages on the world-wide web and whose edges correspond to hyperlinks between them. In [9–11] experiments on the local and global properties of the www graph were carried out using www crawls. The main result obtained was that the in- and out-degree distributions are both power law with different exponents. Additionally a number of other features of the directed graph were identified. These included

(i) a maximal mutually reachable subset of nodes known as a strongly connected component (SCC);

(ii) an IN region containing nodes not in the SCC from which there are paths into the SCC;

(iii) an OUT region formed from nodes not in the SCC that can be reached from the SCC;

(iv) tendrils formed from nodes not in the SCC, IN or OUT which are connected out of IN or in to OUT.

In addition there are a small number of disconnected components. The SCC contains about 56 million nodes, and the IN and OUT regions, and the tendrils, contain about 44 million nodes each [11]. The strongly connected component is sometimes called the *giant component*. The relative sizes of these regions are very different from those that would be predicted using Poisson connectivity statistics. This suggests that models to explain this structure must include non-Poisson statistics as a key indegredient, or generate such statistics as part of their kinetics.

In [12], a model of a directed growing random graph in which nodes with high degree were more likely to gain new edges was introduced and solved. The distributions of in- and out-degree were both found to be power law, with, in general, different exponents. By adjusting the parameters in the model, the model was found to match the exponents in the power laws and the mean connectivity of the www [11].

In this paper we study some more detailed global and local properties of the graph introduced in [12], with a view to understanding other properties of the www graph identified in [11]. A similar approach was recently taken in [13] to an undirected growing graph and in [14] to determine the giant component of generic directed networks. In addition to the www, our model is applicable to other real directed multigraphs with power law degree distributions, such as the call graph [15]. This graph is formed from edges which represent telephone calls made in one day and nodes which are phone numbers that make or receive a call in that day.

This paper is organised as follows. In Section 2 we describe the model in more detail. In Sections 3, 4 and 5 we consider the minimum path on the graph, the nodes with the highest connectivity and the appearance of the giant component, respectively. In Section 6 we examine the number and size of the connected components and in Section 7 we summarise our findings and make suggestions for future work.

II. THE MODEL

In [12], a model for a growing random directed network was introduced which yields a connected multi-graph. At each time step,

(a) with probability p a new node is attached to a node with in-degree i with rate $i + \lambda$ and

(b) with probability $q = 1 - p$ a new directed edge is formed from a node of out-degree

j to a node of in-degree i with rate $(i+\lambda)(j+\mu)$.

The model has three tunable parameters, $0 \le p \le 1$, $\lambda > 0$ and $\mu > -1$. As the network evolves the degree distribution, $N_{ij}(t)$, defined as the average number of nodes with in-degree i and out-degree j at time t, obeys the equation [12]

$$\frac{dN_{ij}(t)}{dt} = \left[\frac{(i-1+\lambda)N_{i-1,j} - (i+\lambda)N_{ij}}{I+\lambda N}\right] + q\left[\frac{(j-1+\mu)N_{i,j-1} - (j+\mu)N_{ij}}{J+\mu N}\right] + p\delta_{i0}\delta_{j1} \quad (1)$$

where $i \ge 0$ and $j \ge 1$. The first term on the right hand side of Eq. (1) represents the sum of the changes of in-degree due to processes (a) and (b), the second term represents the change in out-degree due to process (b) and the final term represents the addition of new nodes. The total number of nodes is $N(t) = \sum_{ij} N_{ij}(t)$, the total in-degree is $I(t) = \sum_{ij} iN_{ij}(t)$ and the total out-degree is $J(t) = \sum_{ij} jN_{ij}(t)$. It is simple to show that $N(t) = pt + 2$ and $I(t) = J(t) = t + 1$, where we assume that initially the network is composed of two nodes connected together. By solving Eq. (1) for a few values of i and j, it is easy to see that in general $N_{ij}(t) = tn_{ij}$ for large t. Hence it is possible to write down a recursive relation for n_{ij}, from which it is possible to show [12] that the in-degree and out-degree distributions are power laws with exponents

$$\nu_{in} = 2 + p\lambda, \qquad \nu_{out} = 1 + q^{-1} + \mu p q^{-1} \quad (2)$$

respectively. The network also displays non-trivial dependence between in- and out-degree, so that the joint probability distribution of in- and out-degree at each site is *not* equal to the product of the distributions of in- and out-degree. It seems likely that this will be the case for both the www and the call graph although to date this question has not been investigated. It is also important to note that the degree distributions on neighbouring nodes are not independent [16]. Again this property is likely to hold for the www but will be much more difficult to measure than the correlation between in- and out-degree on a single node.

In [11] it was shown that the average degree of each node on the www is ≈ 7.5, that $\nu_{in} \approx 2.1$ and $\nu_{out} \approx 2.7$. Using Eq.(2), and noticing that the average degree of each node is $1/p$, gives values for the parameters in the model [12] as

$$p = 0.13, \qquad \lambda = 0.75, \qquad \mu = 3.55. \tag{3}$$

We have performed a simulation of this model with values of the parameters given by Eq.(3). The distribution of in- and out-degree is shown in Figs. 1 and 2 for a graph grown to a size of 107,000 nodes and 800,000 edges.

As predicted [12], these distributions are power laws with exponents given by Eq.(2). These figures are similar to those obtained for the www [11] and the call graph [15] in that the power law for the in-degree is clearly seen over a wider range of degrees than that for the out-degree.

III. MINIMUM PATH

The length of the minimum path between each pair of nodes in the directed graph may be found by considering the adjacency matrix, $A = a_{ij}$, where $a_{ij} = 1$ if there exists a directed edge from node i to node j and $a_{ij} = 0$ otherwise. This method considers increasing powers of A. The elements of A^k show the numbers of paths of length k between each pair of nodes. Thus, the length of the minimum path between node i and node j equals k if the (i, j) element of A^k is positive and the corresponding elements of A, A^2,..,A^{k-1} are all zero. The method is of complexity $O(N^3)$. There are, however, more computationally efficient methods available. For example, Dijkstra's method [17] for finding the minimum path from a specified node to all other nodes is of complexity $O(N^2)$ in its standard form. This method can, however, be implemented with complexity $O(E + NlogN)$ [18], where E is the number of edges. Thus, finding the lengths of the minimum paths between all pairs of nodes is of complexity $O(NE + N^2logN)$. This implementation exploits the sparsity present in many graphs, including the ones generated in this application.

Applying the multiplication method described above to the model in Section II we found that $32 \pm 2\%$ of node pairs had a directed path between them. This is comparable with the www where paths exist between 24% of pairs [11]. A typical distribution of minimum path lengths is shown in Fig. 3. This is very similar to that obtained in [19] for the www graph.

One can make an estimate of the mean minimum path length by the following heuristic argument. Define the mean minimum path length to be l and the mean out-degree within the strongly connected cluster to be m. Then counting up all the nodes we have that

$$m^l \sim N \tag{4}$$

and hence

$$l \sim \frac{\log N}{\log m}. \tag{5}$$

This expression assumes that there are no correlations between the degree distributions at neighbouring nodes and that all pairs of nodes that have a finite path between them are part of the SCC. Consequently this result should be used with some caution. In Figure 4 there is a plot of the mean minimum path length against $\log N$. From the slope of the line one can estimate that the mean out-degree, $m, \approx 50$. This indicates that the degree distribution within the SCC is very different from that of the network as a whole, which has a mean out-degree of 7.5. This also suggests that the contribution of the highly connected nodes to the shortest paths is very high.

In [20] a similar, though more sophisticated, argument was used to obtain an expression for the mean minimum path length as a function of the system size and the mean number of first and second nearest neighbours. Applied to our data, the expression in [20] yields similarly high values for the mean connectivities, for the reasons given above.

IV. HIGHLY CONNECTED NODES

It has been established that high connectivity nodes play an important role as hubs in communication and networking. This fact is being exploited to design and build efficient search and crawl algorithms. In [4] a number of local search strategies were introduced for power law networks which utilised the high connectivity of some nodes. These strategies yielded search times that scaled sub-linearly with the size of the graph. The analysis in [4], and that in an earlier work on undirected graphs [15], used a power law degree distribution with a cut-off in the large degree limit. This was done by introducing by hand a single node with the largest degree, which scaled as a power of N.

In our model it is much more natural to allow the degree of the most highly connected nodes to emerge naturally. Imagine that the degree distribution of a network with N nodes is power law, with exponent $\nu > 2$. Then the *average* degree of the most highly connected node, m, is given by

$$\sum_{i=m}^{\infty} i^{-\nu} \sim \frac{1}{N} \tag{6}$$

which in the limit of large N, and hence m, yields

$$m \sim N^{\frac{1}{\nu-1}}. \tag{7}$$

By using the values of the power laws for the www, we find that the largest in-degree m_{in} and the largest out-degree m_{out} behave as

$$m_{in} \sim N^{0.91}, \quad m_{out} \sim N^{0.59}. \tag{8}$$

Thes results were verified numerically, as shown in Fig. 5.

These predictions are a direct consequence of the power-law distributions of in- and out-degree. Applying them to one recent study of the www [11], (where $N = 2 \times 10^8$), suggests there was a node visited on that crawl of the www with an in-degree of about 110 million hyperlinks and another node (probably different) with an out-degree of 780 thousand hyperlinks. These values appear to be plausible, since pages with out-degree greater than this are easy to find on the www. For instance, in July 2001, searching on the word "Netscape" in the Altavista search engine revealed a results page with an out-degree of over 8 million.

V. GIANT COMPONENT

The appearance of a giant strongly connected component in this system is analogous to the formation of a gel in an aggregation process. This transition is signalled by the divergence of some moment of the size distribution, (see for instance [21]). Consequently, we expect the formation of a giant component to be signalled by the divergence of a moment of $N_{ij}(t)$. The lowest non-trivial moment of $N_{ij}(t)$ is $M(t)$, defined by

$$M(t) = \sum_{ij} ij N_{ij}(t). \tag{9}$$

By multiplying Eq.(1) by ij and summing over i and j we can show that $M(t)$ obeys the differential equation

$$\frac{dM(t)}{dt} = M(t)\left[\frac{1}{I+\lambda N} + \frac{q}{J+\mu N}\right] + \frac{\lambda J}{I+\lambda N} + \frac{q\mu I}{J+\mu N}. \tag{10}$$

It is simple to show that for large times $M(t) \sim t^{\alpha}$ where $\alpha = max(\beta, 1)$ and

$$\beta = \frac{1}{1+p\lambda} + \frac{q}{1+p\mu}. \tag{11}$$

Hence $\beta = 1$ is a transition point in the behaviour of $M(t)$. When $\beta < 1$ the average number of routes through a node, $(= M(t)/N(t))$, is finite whereas for $\beta \geq 1$ it diverges as $t \to \infty$.

Furthermore, in [20] it was shown that the condition for the existence of a giant component is

$$\sum_{ij}(2ij - i - j)N_{ij} > 0. \tag{12}$$

This result holds when the degrees at neighbouring nodes are uncorrelated, a property that does not hold for growing random systems, [16]. However, it is interesting to examine the behaviour of this summation for our model. The second and third terms in this summation are the total in- and out-degree $I(t)$ and $J(t)$. These grow linearly with time. Hence when $\beta > 1$ the summation is dominated by the first term as $t \to \infty$ and the condition is satisfied.

Consequently there is strong evidence that $\beta > 1$ is a *sufficient condition* for the existence of a giant component *in the limit* $t \to \infty$. For the values of the parameters in Eq.(3), $\beta \approx 1.51$ and a giant component is expected in the www. As far as we are aware, there is not enough data about in the published literature to allow a similar analysis for call graphs.

VI. COMPONENT DISTRIBUTION

We used Tarjan's algorithm [22] to find the strong components, or maximal subsets of nodes which are mutually reachable, for a number of graphs generated using the procedure outlined in Section II. This algorithm creates a depth first spanning forest in which the vertices of each strong component correspond to the vertices of a single sub-tree of this forest. The complexity of the algorithm is linear in terms of the number of edges and the number of vertices of the graph. We found that the graphs all had one large strongly connected component, as predicted in the previous section. In Fig. 6 there is a graph of the log of the size of the strongly connected component against log N. This figure was obtained by finding the size of the giant component of 1 graph with 10^6 edges, 10 graphs of 10^5 edges, etc..., down to 10^4 graphs with 100 edges. The position of the intercept of the y-axis, and the fact that the graph has slope 1.00 ± 0.02 suggests that the giant component is of size $(0.32 \pm 0.02)N$. This compares well with the www [11], where around 28% of the sites are within the strongly connected component.

Of the graphs we studied, 97% had no strong components other than the giant component. Of the 3% that had more than one component, the size of the smaller components contained of order 1 nodes. Consequently, the topology of these graphs is made up of a single strongly connected giant component, no OUT region and an IN region made up of a forest of rooted trees directed towards the giant component, which contain either 0 or $O(1)$ nodes in strongly connected components. The absence of an OUT region is caused by the particular choice of kinetics for this model. New nodes are always connected to another node and have out-degree 1 and in-degree 0. Conversely, no new nodes are ever added with in-degree 1 and out-degree 0, which would provide the seed for an OUT region.

VII. DISCUSSION

We have studied a number of the properties of a network introduced in [12] and compared them with empirical studies of the www. We were also able to derive a criterion for the appearance of a giant component in directed growing networks of this type. The distribution of the minumum path lengths, the connectivity of the hub sites and the size of the giant component are all in good quantitative agreement with the www. The model does less well in capturing the detailed topology of the IN and OUT regions. One imagines that on the www these are both made up of a forest of directed trees in which a number of connected components are embedded. No empirical information is available about the size or frequency of connected clusters within the IN and OUT regions. The graph introduced in [12] has no OUT component, but the IN component is qualitatively the same as the www.

There are a number of avenues for further study. As has been pointed out elsewhere, there is a general need for empirical results from the www or call graphs on the joint probability distribution of in- and out-degree. If this were known then models of this type could be critically evaluated and refined with greater precision. In addition, using models of random growing directed networks, it would be interesting to investigate which growth rules give rise to realistic non-trivial IN and OUT regions as well as a giant component, and to characterise the size distributions of the trees, and the size distribution of the connected components, within the IN and OUT regions.

ACKNOWLEDGMENTS

We would like to thank the EPSRC and the Leverhulme Trust for their financial support and Steve Noble and René D'Hulst for useful discussions.

FIGURES

FIG. 1. Log of frequency against log of in-degree for a graph of 107,000 nodes and 800,000 edges.

FIG. 2. Log of frequency against log of out-degree for a graph of 107,000 nodes and 800,000 edges.

FIG. 3. Frequency against minimum path length for a 40 graphs with 5000 edges and an average of 671 nodes.

FIG. 4. Mean minimum path length against log N for 40 realisations of each random graph.

FIG. 5. Log of largest degree against log N averaged over 20 realisations of each graph. The upper line is the in-degree and the lower the out-degree. The straight lines are given by Eq. (8).

FIG. 6. Log of the size of the giant component against log N

REFERENCES

[1] R. Albert and A.-L. Barabási, *cond-mat/0106096.*

[2] S. N. Dorogovtsev and J. F. F. Mendes, *cond-mat/0106144.*

[3] S. Janson, T. Luczak and A. Rucinski, *Random Graphs* (Wiley, New York, 2000).

[4] L.A. Adamic, R. M. Lukose, A.R. Puniyani and B. A. Huberman, *n.lin/0103016.*

[5] R.Cohen, K. Erez, D. ben-Avraham and S. Havlin, Phys. Rev. Lett. **85** 4626 (2000).

[6] D. S. Callaway, M. E. J. Newman, S.H. Strogatz and D. J. Watts, Phys. Rev. Lett. **85** 5468 (2000).

[7] B. Hayes, American Scientist **88** Nos. 1 and 2 (2000).

[8] S. H. Strogatz, Nature **410** 268 (2001).

[9] A.-L. Barabási and R. Albert, Science **286** 509 (1999).

[10] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, Trawling the Web for cyber communities, *Proc. 8th WWW Conference* (1999).

[11] A. Broder, R. Kumar, F. Maghul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, Computer Networks **33** 309 (2000).

[12] P. L. Krapivsky, G. J. Rodgers and S. Redner, Phys. Rev. Lett. **86** 5401 (2001).

[13] D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman and S. H. Strogatz, *cond-mat/0104546.*

[14] S. N. Dorogovtsev, J. F. F. Mendes and A. N. Samukhin, *cond-mat/0103629.*

[15] W. Aiello, F. Chung and L.Lu, A random graph model for massive graphs, *STOC '00, Proceedings of the thirty-second annual acm symposium on Theory of computing* pp. 171-180 (2000).

[16] P. L. Krapivsky and S. Redner, Phys. Rev E **63** 066123 (2001).

[17] E. Dijkstra, Numeriche Mathematics **1** 269 (1959).

[18] M. L. Fredman and R.E. Tarjan, Journal of the ACM, **34** 596 (1984).

[19] L.A. Adamic, The Small World Web, Lect. Notes Comput. Sci. **1696** 443 (1999).

[20] M. E. J. Newman, S. H. Strogatz and D. J. Watts, to be published in Phys. Rev. E., *cond-mat/0007235*.

[21] R. M. Ziff, M. H. Ernst and E. M. Hendriks, J. Phys. A. **16** 2293 (1983).

[22] R. E. Tarjan, SIAM J. Comput. **1** 146 (1972).