Limits and power laws of models for the web graph and other networked information spaces

Anthony Bonato[1] and Jeannette Janssen[2] *

[1] Department of Mathematics
Wilfrid Laurier University
Waterloo, ON
Canada, N2L 3C5
abonato@rogers.com
[2] Department of Mathematics and Statistics
Dalhousie University
Halifax, NS
Canada, B3H 3J5
janssen@mathstat.dal.ca

Abstract. We consider a generalized copy model of the web graph and related networks, and analyze its limiting behaviour. The model is motivated by previously proposed copying models of the web graph, where new nodes copy the link structure of existing nodes, and a certain number of additional random links are introduced. Our model parametrizes the number of random links, thereby allowing for the analysis of threshold behaviour. We consider infinite limits of graphs generated by our model, and compare properties of these limits with orientations of the infinite random graph. As well, we analyze the power law behaviour of the in-degree distribution of graphs generated by our model.

1 Introduction to the model

The overwhelming success of search engines that use a graph-based ranking system (Google being the most famous example) has made the *Web graph* a popular object of study. The Web graph is the graph formed by web pages or sites as nodes, and hyperlinks as directed edges. The Web graph is the most famous, and most complex, example of a *Networked Information Space (NIS).* A NIS is a collection of information containing entities which are linked together. Other

* The authors gratefully acknowledges support from NSERC research grants, and from a MITACS grant. The first author acknowledges support from a Wilfrid Laurier University Senior Research Fellowship.

examples are digital libraries, which consist of publications linked by references [10], or networked databases recording phone calls or financial transactions.

In order to exploit the link structure of Networked Information Spaces, for ranking of search results, clustering, or focused crawling for example, we need to understand its generative process. Typically, the graphs in question have the property that their in-degree distributions satisfy a *power law*: for a positive integer k, the proportion of nodes of in-degree k is approximately $k^{-\gamma}$, where γ is an exponent that is generally observed to be between 2 and 3. This qualifies them as so-called *scale-free networks*. Several stochastic models for the generation of scale-free networks have been proposed. In the *copying models* of [1, 11] a new node copies part of the link environment of an existing node, while adding an additional number of random links. Copy models seem to be especially appropriate to model the link structure of a NIS. Namely, it is very likely that a new information entity (for example a web page or paper), is modelled on an existing one, and hence, will create a link environment that will have a large overlap with that of its model, but will also include some new links. Other recent models use other paradigms (such as preferential attachment) for link creation; see [2–4]. A recent generation model for the internet based on *Heuristically Optimized Trade-Offs* [12] implicitly contains the concept of copying.

We continue our study of [5, 6] of the *infinite* limits of graphs generated by stochastic models. One important reason why we consider such limits is to investigate the consequences of the choices made in the design of the model. In a certain sense, the infinite limit magnifies the properties of the finite graphs that lead to it. In our previous work, we considered generation processes for undirected graphs, and their measure of convergence to the *infinite random graph*, or R (introduced by Erdős and Rényi in [9]). For a fixed positive integer, and real number $p \in (0, 1)$, the graphs in $G(n, p)$ possess n nodes, and every pair of distinct nodes is joined independently and with probability p. If we consider the limit of the graphs in $G(n, p)$ as n tends to infinity, then the resulting graph will with probability 1 be isomorphic to R. As noted in [6], convergence to R may be viewed as loosing all structure, comparable in a sense to convergence to white noise.

The graph R is undirected, as are the limiting graphs introduced in [5]. However, the link structure of a NIS often is by nature a directed graph: for a web page, it is easy to see what pages it points to, but very hard to find out what pages point to it. For scientific papers, it is definitely not the same to cite or be cited. Hence, the question arises as to what infinite graphs could be taken as directed versions of R. A suitable "generic" infinite directed graph is the *Acyclic Random Oriented Graph*, or ARO, introduced and investigated recently in [8]. ARO is an orientation of R. Like R, ARO is uniquely defined (up to isomorphism) by a certain adjacency property (see Section 2), is the limit of a sequence of finite acyclic directed graphs, and contains every acyclic directed graph as an induced subgraph.

However, our research shows that graphs generated by the copy models are unlikely to converge to ARO. As we prove in Theorem 3, the limits are *isomorphic* to ARO with low probability, but in some cases are *homomorphically equivalent* with ARO with high probability. In order to classify the limits that we do obtain, we introduce the concept of a *n*ear-ARO. There are infinitely many non-isomorphic near-ARO's; this could imply a certain sensitivity of the copying model to initial conditions. Near-ARO's have a rich structure; see Theorem 1. While convergence to ARO may be viewed as a process of loosing all structure, convergence to a near-ARO may be viewed as loosing a *significant amount* of structure.

We introduce a model $D(p, \rho, H)$ motivated by the copying models. Like the copying models, our model uses copying of the existing link structure as its basic rule for link creation. Our model allows for variation of the number of random links. The three parameters of this model are a fixed *copying probability* $p \in (0, 1)$, a *random link function* $\rho : \mathbb{N} \to \mathbb{R}$, defined by

$$\rho(t) = \alpha t^s,$$

where α and s are non-negative real constants so that $s \in [0, 1]$, and a fixed finite acyclic *initial digraph* H.

1. At $t = 0$, set $G_0 = H$.
2. For a fixed $t \geq 0$, assume that G_t has been defined, is finite, and contains G_0 as an induced subdigraph. To form G_{t+1}, add a new node v_{t+1} to G_t and choose its out-neighbours as follows.

(a) Choose an existing node u from G_t uniformly at random (u.a.r.). The node u is called the *copying node.*

(b) For each out-neighbour w of u, independently add a directed edge from v_{t+1} to w with probability p. In addition, choose $\lfloor \rho(t) \rfloor$-many (not necessarily distinct) nodes from $V(G_t)$ u.a.r., and add directed edges from v_{t+1} to each of these nodes. The latter edges are called *random links.*

(c) Make the digraph G_{t+1} simple by removing any parallel edges.

At each time-step t, our model adds approximately $\rho(t)$-many random links between the new node and the existing nodes. Theorem 3 shows that if $\rho(t) \in \theta(t)$, then the limit is a near-ARO with high probability. As s tends to 1, with high probability the limit, while not a near-ARO, behaves increasingly like a near-ARO. On the other hand, Theorem 5 shows that we loose power law behaviour if we have more than a constant number of random links. Hence, an interesting "grey area" where $0 < s < 1$ emerges; we will elaborate further on this in Section 4.

2 Limits and the $D(p, \rho, H)$ models

Before we state the main results for this section, we require a few definitions. If u is a node in a digraph G, then let

$$N_{\uparrow}(u) = \{v \in V(G) : (u, v) \in E(G)\}$$

be the *out-neighbourhood* of u in G. If $(G_t : t \in \omega)$ is a sequence of digraphs with G_t an induced subdigraph of G_{t+1}, then define the *limit* of the G_t, written

$$G = \lim_{t \to \infty} G_t,$$

by

$$V(G) = \bigcup_{t \in \mathbb{N}} V(G_t),$$

and

$$E(G) = \bigcup_{t \in \mathbb{N}} E(G_t).$$

A digraph G is *good* if G is acyclic, has no infinite directed paths, and each node of G has finite out-degree. For example, if G_t is generated by our model $D(p, \rho, H)$, then any limit $G = \lim_{t\to\infty} G_t$ is good. A digraph G is an *acyclic random oriented graph*, or *ARO* for short, if G is good, and for each finite set $S \subset V(G)$, there are infinitely many nodes u such that $S = N_{\uparrow}(u)$. AROs were introduced and investigated recently in [8], where it was proved (among other things) that a countable ARO is unique up to isomorphism. Hence, we will refer to this unique isomorphism-type simply as ARO. (Strictly speaking, we are using the *inverse* of ARO as defined in [8], where the inverse of a digraph results by reversing the orientations of all the directed edges. Since we have only have use for ARO as defined above, we will keep our notation.) As noted first in [8], ARO results from a suitable orientation of the infinite random graph R. Indeed, ARO may be defined probabilistically: let $\mathbb{N}$ be the set of nodes, allow all edges to be directed backward (that is, (i, j) is a directed edge only if $j < i$), and adopt these edges independently with probability 2^{i+j}. The digraph ARO results with probability 1 from this random digraph.

We say that a digraph G is a *near-ARO* if it is good, and for each finite set $S \subset V(G)$, there is a node $u \in V(G)$ such that $S \subseteq N_{\uparrow}(u)$. ARO is clearly near-ARO. However, there are many examples of countable near-ARO digraphs that are not isomorphic to ARO; see Corollary 1 below.

We say that an undirected graph G is *algebraically closed*, or a.c. for short, if for each finite subset U of nodes of G, there is a node $z \in V(G)\backslash U$ such that z is joined to each node of U. For example, an infinite clique and R are examples of a.c. graphs. A *homomorphism* from the digraph G to H is an edge-preserving mapping from $V(G)$ to $V(H)$. The digraphs G and H are *homomorphically equivalent*, written $G \leftrightarrow H$, if there is a homomorphism from G to H and from H to G. Note that isomorphic digraphs are homomorphically equivalent, although the converse fails.

The following theorem (whose proof relies on König's infinity lemma and the back-and-forth method, and so is omitted) characterizes near-ARO digraphs up to homomorphic equivalence.

Theorem 1. *Let $G = \lim_{t\to\infty} G_t$ be a good digraph. The following are equivalent.*

1. *The underlying graph of G (formed by forgetting the orientation of each directed edge) is a.c.*
2. *The digraph G is a near-ARO.*
3. *The digraph $G \leftrightarrow$ARO.*
4. *For all countable good digraphs H, H admits a homomorphism into G.*

While the digraph ARO is unique up to isomorphism, the following corollary demonstrates that the maximum possible number of non-isomorphic near-ARO digraphs exist. We write $2^{\aleph_0}$ for the cardinality of the set of real numbers.

Corollary 1. *There are $2^{\aleph_0}$ many non-isomorphic countable near-ARO digraphs.*

We say that a digraph satisfies the *locally near-ARO* adjacency property if it is good, and for all finite sets of nodes S that are in the out-neighbourhood of some other node y, there is a node whose out-neighbours include S. Clearly, a near-ARO digraph is locally near-ARO; Theorem 3 (4) will demonstrate that the converse is false. Our next result shows that for all values of s, limits of graphs generated by $D(p, \rho, H)$ are locally near-ARO with high probability.

Theorem 2. *Fix $p \in (0, 1)$, ρ, and H. With probability 1, the limit*

$$G = \lim_{t\to\infty} G_t$$

of graphs generated by the model $D(p, \rho, H)$ is locally near-ARO.

Proof. Since a countable union of measure 0 subsets has measure 0, it suffices to show that for a fixed $y \in V(G)$ and a finite $S \subseteq N_{\uparrow}(y)$ the probability that there is no node joined to all of S is 0 (since there only countably many choices for y and S in G).

Let t_0 be the least integer such that y and S are in $V(G_{t_0})$. Let $|V(G_{t_0})| = m$ and $|S| = i$. If $t \geq t_0$, the probability that y is chosen as copying node in G_t equals $\frac{1}{m+t-t_0}$. Given that y is the copying node,

v_t is joined to all of S with probability p^i. Then the probability that no node of G is joined to all of S is at most

$$\prod_{t=t_0}^{\infty}\left(1-\left(\frac{1}{m+t-t_0}\right)p^i\right)=0. \quad \Box$$

Our main result is the following theorem, which demonstrates that as s tends to 1, graphs G generated by $D(p,\rho,H)$ share more and more properties of a near-ARO. Further, the graphs G are very rarely isomorphic to ARO. For a positive integer n, we say that a digraph G is *n-near-ARO* if it is good, and for each set $S \subset V(G)$ of cardinality n, there is a node $u \in V(G)$ such that $S \subseteq N_{\uparrow}(u)$. Observe that a digraph is near-ARO if and only if it is n-near-ARO for all positive integers n.

Theorem 3. *Fix $p \in (0,1)$, $\rho = \alpha t^s$, and H. Let G be the limit of a sequence of digraphs generated according to the model $D(p,\rho,H)$.*

1. *If $s = 1$, then with probability 1 G is near-ARO.*
2. *If $s \in [0,1)$, then with probability 1 G is $\lfloor\frac{1}{1-s}\rfloor$-near-ARO.*
3. *If $s \in [0,1)$, then with positive probability G is not near-ARO.*
4. *For all $s \in [0,1]$, with probability 1 G is not isomorphic to ARO.*

Theorem 3 suggests a *threshold* behaviour for convergence to a near-ARO: as s tends to 1, with high probability the limit G acquires more and more properties of a near-ARO, but with positive probability is not near-ARO. At $s = 1$, we obtain a near-ARO with high probability.

Proof of Theorem 3. We sketch a proof of (2) only. Let $G = \lim_{t\to\infty} G_t$. It is straightforward to see that G is good. As in the proof of Theorem 2, it suffices to show that for a fixed finite $S \subseteq V(G)$ the probability that there is no node joined to all of S is 0.

The proof rests on the following lower bound. For each set S, for each $t \geq t_0$ where t_0 is such that $S \subseteq V(G_{t_0})$, the probability that v_t is joined to every node of S is at least

$$\beta t^{(s-1)|S|}(1+o(1)) \tag{1}$$

where $\beta \in (0,1)$ is a constant that does not depend on t. The proof of this bound uses induction on the size of S.

Fix S a set of nodes with $|S| \leq \lfloor \frac{1}{1-s} \rfloor$. Then $t^{(s-1)|S|} \geq t^{-1}$. Hence, by (1), the probability that there is no node of G joined to every node of S is at most

$$\prod_{t=t_0}^{\infty} \left(1 - \beta t^{-1}(1 + o(1))\right) = 0. \quad \square$$

Theorem 3 (1) may be generalized to other values of $\rho(t)$ (which are not necessarily a power of t) with only minor changes in the proof.

Theorem 4. *Let G be the limit of a sequence of graphs generated according to the model $D(p, \rho, H)$, with ρ a non-negative, monotone increasing function $\rho : \mathbb{N} \rightarrow \mathbb{R}$ satisfying the condition:*

$$\sum_{t=0}^{\infty} \frac{\rho(t)}{t^2} = \infty. \tag{2}$$

Then with probability 1, G is near-ARO.

3 Degree distributions of the $D(p, \rho, H)$ models

When do the models $D(p, \rho, H)$ produce digraphs whose in-degree distributions follow power laws? We find in the following results that power laws are sensitive to the choice of ρ.

Theorem 5. *Fix $p \in (0, 1)$ and H. Let G be the limit of a sequence of graphs generated according to the model $D(p, \rho, H)$, where $\rho(t) = \alpha t^s$. Then the degree distribution of G_t converges to a power law distribution if and only if $s = 0$.*

Proof. Let $X_i(t)$ be the expected number of nodes of in-degree i at time t. Suppose that

$$\lim_{t \to \infty} \frac{X_i(t)}{t} = b_i = ci^{-d},$$

for some positive constants c and d. It follows that

$$b_i = p((i-1)b_{i-1} - ib_i) + \lfloor \rho(t) \rfloor (b_{i-1} - b_i) + o(1).$$

By definition b_i is a constant, so either $b_i - b_{i-1} = 0$ or $\rho(t) = \alpha t^0 = \alpha$. If $b_i = b_{i-1}$, then this contradicts that $b_i = ci^{-d}$. Therefore, $s = 0$ and $\rho(t) = \alpha$. We omit the details that if $\rho(t) = \alpha$, then a power law is obtained. □

4 Conclusions and future work

We introduced a new model $D(p, \rho, H)$ of the web graph and other Networked Information Spaces, motivated by the copying models of the web graph proposed by [1, 11]. $D(p, \rho, H)$ provides a continuum of models, whose structural properties depend largely on the number of random links parameterized by $\rho = \alpha t^s$.

We have seen that for all values of $s \in [0, 1]$, our model generates limit graphs G which satisfy the locally-near-ARO adjacency property. As s tends to 1, G becomes increasingly random, until at $s = 1$ it is with probability 1 homomorphically equivalent with a certain random acyclic digraph, ARO. Hence, on the one hand, the model $D(p, \rho, H)$ is robust: a large number of random links must be added at each time-step to ensure a random-like structure. Further, the choices of p and H seem to have little impact on the structure of the limit. On the other hand, we obtain power laws only if there are at most a constant number of random links. Hence, for any choice of ρ with $0 < s < 1$, there is an interesting "grey area" that emerges: the limits are not completely random, nor do we obtain power laws. We do not understand exactly the in-degree distributions that arise when $s \in (0, 1)$. We plan on analyzing these distributions in future work.

References

1. M. Adler, M. Mitzenmacher, Towards compressing web graphs, In: *Poceedings of the IEEE Data Compression Conference (DCC)*, 2001.
2. W. Aiello, F. Chung, L. Lu, Random evolution in massive graphs, In: 42nd IEEE Symposium on Foundations of Computer Science, 2001.
3. R. Albert and A. Barabási, Emergence of scaling in random networks, *Science* **286** (1999) 509-512.
4. B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, The degree sequence of a scale-free random graph process, *Random Structures Algorithms* **18** (2001) 279-290.

5. A. Bonato and J. Janssen, Infinite limits of copying models of the web graph, accepted in *Internet Mathematics*.
6. A. Bonato and J. Janssen, Limits and power laws of web graph and biological network models, submitted.
7. P.J. Cameron, The random graph, in: *Algorithms and Combinatorics* **14** (R.L. Graham and J. Nešetřil, eds.), Springer Verlag, New York (1997) 333-351.
8. R. Diestel, I. Leader, A. Scott, and S. Thomassé, Partitions and orientations of the Rado graph, submitted.
9. P. Erdős and A. Rényi, Asymmetric graphs, *Acta Math. Acad. Sci. Hungar.* **14** (1963) 295-315.
10. Y. An, J. Janssen, E. Milios, Characterizing the citation graph of computer science literature, accepted in *Knowledge and Inf. Systems (KAIS)*.
11. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, Stochastic models for the web graph, In: *Proceedings of the 41th IEEE Symp. on Foundations of Computer Science*, 2000.
12. A. Fabrikant, E. Koutsoupias and C. Papadimitriou, Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet, In: *Proceedings of the 34th Symposium on Theory of Computing*, 2002.