A survey of models of the web graph

Anthony Bonato*

Department of Mathematics
Wilfrid Laurier University
Waterloo, ON
Canada, N2L 3C5
abonato@rogers.com

Abstract. The web graph has been the focus of much recent attention, with several stochastic models proposed to account for its various properties. A survey of these models is presented, focussing on the models which have been defined and analyzed rigorously.

1 Introduction

The web graph W has nodes representing web pages, and edges representing the links between pages. The graph W is massive: at the time of writing, it contains billions of nodes and edges. In addition, W is dynamic or *evolving*, with nodes and edges appearing and disappearing over time. The explosive growth of W itself is mirrored by the recent rapid increase in research on structural properties, stochastic models, and mining of the web graph. There are now several survey articles [10, 19, 44], and several books on W [6, 22, 31] (including a popular book by Barabási [7]). A new mathematics journal devoted to research related to W, *Internet Mathematics,* was recently launched.

The purpose of the present survey is to highlight recent stochastic models which are used to model W. We focus on six of these models, chosen both for their various design elements, and because they have been rigorously defined and analyzed. For more on information on web mining and the mathematics of web search engines, the reader is directed to Chakrabarti [22] and Henzinger [39].

For background on graph theory and random graphs, the reader is directed to [8, 9, 30, 40]. We use the notation $\mathbb{N}$ for the nonnegative integers and $\mathbb{N}^+$ for the positive integers. If A is an event in a

* The author gratefully acknowledges the support from an NSERC Discovery grant and from a MITACS grant.

probability space Ω, then $\mathbb{P}(A)$ is the probability of the event in the space Ω; if X is a random variable with domain Ω, then $\mathbb{E}(X)$ is the expectation of X. All logarithms are in base 2.

2 Properties of W

We begin with a brief overview of some of the experimental data and structural properties observed in W collected from various web crawls. The overview will highlight experimental features of W that models of the web graph often attempt to replicate. For a more detailed introduction to experimental data on W, the reader is directed to Chapter 3 of Dorogovtsev, Mendes [31] and Kumar et al. [44].

Arguably the most important properties observed in W are power-law degree distributions. Given an undirected graph G and a non-negative integer k, we define the rational number $P_G(k)$ as follows.

$$P_G(k) = \frac{|\{x \in V(G) : \deg_G(x) = k\}|}{|V(G)|}.$$

In other words, $P_G(k)$ is the proportion of nodes of degree k in G. We will suppress the subscript G if it is clear from context. We say that the degree distribution of G follows a *power law* if for each degree k,

$$P(k) \sim ck^{-\beta},$$

for real constants $c > 0$ and $\beta > 0$. Such distributions are sometimes called *heavy-tailed distributions*, since the real-valued function $f(k) = ck^{-\beta}$ exhibits a polynomial decay to 0 as k tends to ∞. Real-world graphs like W with power law degree distributions are sometimes called *scale-free*. The graph W may be viewed as either a directed or undirected graph. If G is directed, then we may discuss power laws for the in- and out-degree distributions by defining the proportions $P_{in,G}(k)$ and $P_{out,G}(k)$, respectively, in the obvious way.

Based on their crawl of the domain of Notre Dame University, Indiana, Albert et al. [4] claimed that the web graph exhibits a power law in-degree distribution, with $\beta = 2.1$. This claim was supported by an independent larger crawl of the entire web reported in Broder et al. [17], who also found $\beta = 2.1$. There is some evidence in both studies that the out-degree distribution follows a power law with

[4] reporting $\beta = 2.45$ and [17] reporting $\beta = 2.7$. The presence of power law degree distributions reflects a certain *undemocratic* aspect of W: while most pages have few links, a few have a large number. This is perhaps not surprising, since the choice of links from new pages to existing ones is presumably governed by the users own personal or commercial interests. It is interesting to note that power law degree distributions are now known to be pervasive in a variety of real world networks where some degree of choice is involved, such as the telephone call network, the e-mail network, or the scientific citation network; see [31] for other similar networks. Power law degree distributions are also prevalent in biological networks (such as the network of protein-protein interactions in a cell), where evolution is the dominant decision making force in the generation of nodes and edges; see Chung et al. [23]. Models for power law behaviour have long been studied in such disciplines as biology and economics; see Mitzenmacher [45].

So-called *small world graphs* were first introduced by Strogatz, Watts [46] in their study of social networks. One important feature of small world networks is the presence of "short" paths between nodes. To be more precise, define the *distance* from u to v in a graph G, written $d(u, v)$, to be the number of edges in a shortest path connecting u to v, or ∞ otherwise. Define

$$L(G) = \sum_{\{u,v\} \in S} \frac{d(u, v)}{|S|},$$

where S is the set of pairs of distinct nodes u, v of G with the property that $d(u, v)$ is finite. The rational number $L(G)$ is the *average distance* of G. The directed analogue of this parameter, where distance refers to shortest *directed* paths, is denoted $L_d(G)$. The small world property demands that $L(G)$ (or $L_d(G)$ if G is directed) must be much smaller than the order of the graph; for example, $L(G) \in O(\log(|V(G)|))$. As evidence of the small world property for W, in Albert et al. [4] it was reported that $L_d(W) = 19$, while Broder et al. [17] reported $L_d(W) = 16$ and $L(W) = 6.8$. Another measure of global distances in a graph is the *diameter* of G, written $diam(G)$, which is the maximum of $d(u, v)$ taken over all pairs of distinct nodes u and v in G. In contrast with the just cited results on average distance, data from [17] suggests that $diam(W) > 900$.

The web contains many *communities*: sets of pages sharing a common interest. An idea presented in Kleinberg et al. [41] and Kumar et al. [44] is that communities in the web are characterized by dense directed bipartite subgraphs. A *bipartite core* is a directed graph which contains at least one directed bipartite clique as a subgraph, where the directed edges in the subgraph all have terminal nodes of one fixed colour. In their study of communities in W, the authors in [41, 44] show the presence of many more small bipartite cores in W than a directed random graph with the same number of nodes and edges.

3 Models of W

A large number of models for the web graph have been proposed. Such models are useful for several reasons. They deepen our understanding of the generative mechanisms driving the evolution of W. They provide insight into superficially unrelated properties observed in the web. Perhaps most importantly from the point of view of applications, they may aid in the development of the next generation of link-analytic search engines. As discussed in the survey by Bollobás, Riordan [10], the majority of the analysis of models of the web has been heuristic and non-rigorous. A small but growing number of rigorous studies of web graph models have been appearing in the literature, and it is these models that we focus on in the present survey.

Pioneering work on random graphs was first done by Erdos and Rényi [34, 35]. In what is now sometimes called the Erdos-Rényi (ER) model, written $G(n,p)$, we are given n nodes and a fixed real number $p \in (0,1)$. For each of the $\binom{n}{2}$ many distinct pairs of nodes, add an edge between them independently with probability p. In many contexts, p is a function of n, and properties of $G(n,p)$ are studied asymptotically as n tends to ∞. The probability space $G(n,p)$ is often referred to as *a random graph of order* n (an accepted misnomer). Random graphs have been intensively researched, and the subject has spawned several thousand research articles. We direct the interested reader to the texts of Bollobás [9] and Janson et al. [40] for more on the ER model.

The ER model is, in a sense, static or *off-line*: the number of nodes is fixed (although $G(n,p)$ is often viewed as having a variable number of edges with time). In addition, it is straightforward to prove that in $G(n,p)$ the degree of a vertex is binomially distributed. Hence, based on our discussion in Section 1, the ER model is not appropriate as a model of the web graph W. What features would make a good web graph model? The following is a (partial) list of desirable properties that graphs generated by a web graph model should possess, based on the observed properties of W given in the Introduction.

1. *On-line property.* The number of nodes and edges changes with time.
2. *Power law degree distribution.* The degree distribution follows a power law, with an exponent $\beta > 2$.
3. *Small world property.* The average distance (or diameter) is much smaller than the order of the graph.
4. *Many dense bipartite subgraphs.* The number of distinct bipartite cliques or cores is large when compared to a random graph with the same number of nodes and edges.

To aid the reader, we give a chart that summarizes the properties of the various models we will consider. We will focus on the following web graph models, each given an acronym (if they do not already have one) for purposes of comparison: the LCD model of Bollobás et al [14]; the ACL models of Aiello et al [3], the CL model of Chung, Lu [24], [25], the copying model of Kumar et al [43]; the CL-del growth-deletion model of Chung, Lu [27]; and the CFV growth-deletion model of Cooper et al. [29]. A "Y" in the i,j entry of the table means that the model in row i has the property of column j; a "N" or "?" are read similarly. The column "Directed?" refers to whether the model generates directed graphs. The column entitled "β" refers to the possible range of exponent for the power law proven asymptotically for graphs generated by the model, with the value of β dependent on the parameters of the model. If the model produces directed graphs, then the range refers to the in-degree distribution.

Model	Directed?	1	2	3	4	β
LCD	Y	Y	Y	Y	?	3
ACL	Y	Y	Y	?	N	$(2, \infty)$
CL	N	N	Y	Y	?	$(2, \infty)$
copying	Y	Y	Y	?	Y	$(2, \infty)$
CL-del	N	Y	Y	Y	?	$(2, \infty)$
CFV	N	Y	Y	?	?	$(2, \infty)$

As is evident from the chart, all the models in the present survey have property 2, all but the CL model have property 1, while it has yet to be proven that any of them have all four properties. At the time of writing, it is an open problem to find a web graph model that produces graphs which provably has all four properties.

To simplify notation, we present the following general mathematical framework for all the on-line models presented. (Hence, the CL model will not follow this framework.) The model possesses a set of real number parameters, and has a fixed finite graph H as an additional parameter. The model generates by some stochastic process a sequence of finite graphs G_t indexed by $(t : t \in \mathbb{N})$. Unless otherwise stated, for all $t \in \mathbb{N}$, we have that

1. $G_0 \cong H$.
2. G_t is an induced subgraph of G_{t+1};
3. $|V(G_{t+1})| = |V(G_t)| + 1$;
4. $|E(G_t)| \leq |E(G_{t+1})|$.

In all the models we consider, the graphs G_t are defined inductively. In the inductive step, the unique node in $V(G_{t+1}) \backslash V(G_t)$ is referred to as the *new node*, written v_{t+1}, and the nodes of $V(G_t)$ are the *existing nodes*. We refer to a model which generates graphs satisfying all of these conditions as an *evolving graph model*. We note that the choice of H usually has no effect on the value of the power law exponent β, while the choice of real number parameters does generally affect β.

4 Preferential attachment models

The first evolving graph model explicitly designed to model W was given by Albert, Barabási [5]. Informally, the idea behind their model

is a straightforward and intuitively pleasing one: new nodes are more likely to join to existing nodes with high degree. This model is now referred to as an example of a *preferential attachment model.* Albert and Barabási gave a heuristic description and analysis of such a model, and concluded that it generates graphs whose in-degree distribution follows a power law with exponent $\beta = 3$.

The first rigorous attempt to design and analyze a preferential attachment model was given in Bollobás et al. [14]. Their model is called the *Linearized Chord Diagram* or *LCD* model, since an equivalent formulation of the model is via random pairings on a fixed finite sets of integers. The parameter of this model is a positive integer m, where H is a copy of K_1 with a single directed loop. We first describe the model in the case $m = 1$. To form G_{t+1} add a single directed edge from v_{t+1} to v_i, where the node v_i is chosen at random from the existing nodes, with

$$\mathbb{P}(i = s) = \begin{cases} \frac{\deg_{G_{t-1}}(v_s)}{2t-1} & \text{if } 1 \leq s \leq t-1, \\ \frac{1}{2t-1} & \text{if } s = t. \end{cases}$$

This mechanism of joining new nodes to existing ones proportionally by degree we refer to as *preferential attachment.* Observe that the graph G_t is a directed tree for all values of t. Indeed a similar version of this model was previously studied (in a different context) as *random recursive trees*; see [14] for further discussion.

If $m > 1$, define the process $(G_m^t)_{t \geq 0}$ by first generating a sequence $(G_t : t \in \mathbb{N})$ of graphs using the case $m = 1$ on a sequence of nodes $(v_i' : i \in \mathbb{N}^+)$. The graph G_m^t is formed from G_{mt} by identifying the vertices $v_1', v_2', \ldots, v_m'$ to form v_1, identifying $v_{m+1}', v_{m+2}', \ldots, v_{2m}'$ to form v_2, and so on.

Using martingales and the Azuma-Hoeffding inequality (see Theorem 1.19 of [9], for example), Bollobás et al. [14] prove the following theorem.

Theorem 1. *Fix m a positive integer, and fix $\epsilon > 0$. For k a nonnegative integer, define*

$$\alpha_{m,k} = \frac{2m(m+1)}{(k+m)(k+m+1)(k+m+2)}.$$

Then with probability tending to 1 *as* $t \to \infty$, *for all* k *satisfying* $0 \leq k \leq t^{1/15}$,

$$(1-\epsilon)\alpha_{m,k} \leq P_{in,G_m^t}(k) \leq (1+\epsilon)\alpha_{m,k}.$$

Theorem 1 proves that for large t, with high probability the degree distribution of $G_m^{(t)}$ follows a power law with exponent $\beta = 3$ (formally justifying the conclusions derived in [5]). The reader will note that Theorem 1 is stated as a concentration result for degrees in the range $0 \leq k \leq t^{1/15}$; as remarked in [14], this may be extended to degrees $k > t^{1/15}$. The power law exponent $\beta = 3$ is independent of the choice of m.

Bollobás, Riordan [12] prove the following theorem which computes the diameter of G_m^t.

Theorem 2. *Fix an integer* $m \geq 2$ *and a positive real number* ϵ. *With probability* 1 *as* $t \to \infty$, $G_m^{(t)}$ *is connected and*

$$(1-\epsilon)\frac{\log t}{\log\log t} \leq diam(G_m^t) \leq (1+\epsilon)\frac{\log t}{\log\log t}.$$

A set of preferential attachment models different than the LCD model were proposed by Aiello et al. [3]. In [3], four evolving graph models were given. Three models produce directed graphs, while one generates undirected graphs. These models have some advantage over the LCD model, since the power law exponent β may roam over the interval $(2, \infty)$, dependent on the choice of parameters. Not surprisingly, these models are more complex in their description than the LCD model. We summarize only one such model that produces directed graphs, named Model C in [3].

The parameters are $m^{e,e}, m^{e,n}, m^{n,e}, m^{n,n} \in \mathbb{N}^+$ and a fixed finite directed graph H. (We adopt a simpler version of the parameter set in our description; in [3], the numbers are chosen according to some bounded probability distribution.) At time $t+1$, add $m^{e,e}$ directed edges randomly among all nodes. The origins are chosen using preferential attachment with respect to the current out-degree and the destinations are chosen using preferential attachment with respect to the current in-degree. Add $m^{e,n}$ directed edges into v_{t+1} randomly. The origins are chosen using preferential attachment with respect to

the current out-degree. Add $m^{n,e}$ directed edges from v_{t+1} randomly. The destinations are chosen using preferential attachment with respect to the current in-degree. Add $m^{n,n}$ directed loops to v_{t+1}.

The following result was proved in [3] using the Azuma-Hoeffding inequality (see Theorem 3 of [3] for a precise statement of the concentration results). Note that Model C produces directed graphs whose in- *and* out-degree distribution follow power laws.

Theorem 3. *For graphs generated by model C, with probability* 1 *as* t *tends to* ∞*, the out-degree distribution follows a power law with the exponent*

$$\beta = 2 + \frac{m^{n,n} + m^{n,e}}{m^{e,n} + m^{e,e}}.$$

With probability 1 *as* t *tends to* ∞*, the in-degree sequence follows a power law with exponent*

$$\beta = 2 + \frac{m^{n,n} + m^{e,n}}{m^{n,e} + m^{e,e}}.$$

There are other important preferential attachment models such as the model of Cooper and Frieze [28]. Their model is fairly complex, owing to its large number of parameters. Their proof of a power law degree distribution for graphs generated by their model is novel, since it uses martingale techniques along with the Laplace method for the solution of linear difference equations. Both Dorogovtsev et al. [32] and Drinea et al. [33] introduced a variation into preferential attachment where each node is assigned a constant *initial attractiveness* am. The probability that a new node is joined to an existing one u is proportional to its in-degree plus am. Buckley, Osthus [18] gave a rigorous version of this model along the lines of the LCD model. A model using preferential attachment to generate directed graphs in a way different than the ACL and LCD models was given in Bollobás et al. [13].

5 Off-line web graph models

We discuss an interesting off-line model for the web introduced by Chung, Lu [24, 25]. The ER model $G(n, p)$ may be generalized as follows. Let $\mathbf{w} = (w_1, \ldots, w_n)$ be a *graphic* sequence; that is, the

degree sequence of some graph of order n. We define a model for random graphs with expected degree sequence $\mathbf{w}$, written $G(\mathbf{w})$, as follows. The edge between v_i and v_j is chosen independently with probability p_{ij} where p_{ij} is proportional to the product $w_i w_j$. Then $G(n,p)$ may be viewed as a special case of $G(\mathbf{w})$ by taking $\mathbf{w}$ to equal the n-sequence $(pn, pn, \ldots, pn)$. In this way, Chung, Lu [24, 25] consider $G(\mathbf{w})$ where the expected degree sequence is a power law with fixed exponent β in the interval $(2, \infty)$. They refer to such $G \in G(\mathbf{w})$ as *power law random graphs*. The reader will note that the model $G(\mathbf{w})$ generates off-line graphs, unlike all the other models in this survey. The motivation for the study of power law random graphs comes in part from the fact that off-line models are easier to work with mathematically than on-line models. For instance, in contrast to off-line models, for on-line models the probability space for the random graph generated at time-step $t+1$ is different than the one at time-step t.

In [24], the order of connected components of the graphs in $G(\mathbf{w})$ is investigated. The paper [25] proves the following result, which exposes a nice connection between a power law degree distribution and the small world property.

Theorem 4. *Suppose $G \in G(\mathbf{w})$ has n nodes and expected degree sequence $\mathbf{w}$ following a power law with exponent $\beta > 2$. Let G have average degree $d > 1$ and maximum degree m satisfying*

$$\log m \gg \frac{\log n}{\log \log n}.$$

For all values of $\beta > 2$, with probability 1 as n tends to ∞, the graph G is connected with

$$diam(G) = \Theta(\log n).$$

If $2 < \beta < 3$, then with probability 1 as n tends to ∞,

$$L(G) \leq (2 + o(1)) \left(\frac{\log \log n}{\log(1/(\beta - 2))} \right).$$

If $\beta = 3$, then with probability 1 as n tends to ∞,

$$L(G) = \Theta \left(\frac{\log n}{\log \log n} \right).$$

If $\beta > 3$, then with probability 1 as n tends to ∞,

$$L(G) = (1 + o(1)) \frac{\log n}{\log d}.$$

Expected power law degree sequences fall into the more general category of *admissible expected degree sequences* introduced in [25]. The results of Theorem 4 generalize to $G(\mathbf{w})$ with admissible expected degree sequences; see Theorems 1 and 2 of [25].

A recent paper of Chung, Lu [26] uses power law random graphs in the design of a certain off-line model named the *hybrid power law model*. This model generates so-called *hybrid graphs*, whose edge set is the disjoint union of a *global graph* and a *local graph*. The results of [26] show that hybrid graphs satisfy properties 2 and 3 of Section 3, and in addition, are locally highly connected.

6 Copying models

We saw in Section 4 the connection between preferential attachment and power law degree distributions. In this section, we consider an evolving graph model that uses a paradigm different than preferential attachment, but nevertheless with high probability generates graphs with power law degree distributions. The *linear growth copying model* was introduced in Kleinberg et al. [41] and rigorously analyzed in Kumar et al. [43]. It has parameters $p \in (0, 1)$, $d \in \mathbb{N}^+$, and a fixed finite directed graph H with constant out-degree d. Assume G_t has constant out-degree d. At time $t+1$, an existing node, which we refer to as u_t, is chosen u.a.r. from the set of all existing nodes. The node u_t is called the *copying node*. For each of the d out-neighbours w of u_t with probability p, add a directed edge (v_{t+1}, z), where z is chosen u.a.r. from $V(G_t)$, and with the remaining probability $1-p$ add the directed edge (v_{t+1}, w). The authors of [43] use martingales and the Azuma-Hoeffding inequality to prove the following (see Theorems 8 and 9 of [43] for a precise statement of the concentration results).

Theorem 5. *With probability 1 as t tends to ∞, the copying model generates directed graphs G_t whose in-degree distribution converges to a power law with exponent*

$$\beta = \frac{2-p}{1-p}.$$

Property 4 of Section 3, the presence of many dense bipartite subgraphs, is a desirable property for graphs generated by a web graph model. Kumar et al. [43] analyze the model of Aiello et al. [2] (which was defined historically before the ACL models) and demonstrate that this model generates graphs which on average contain few bipartite cliques. Two subgraphs of a graph are *distinct* if they have distinct vertex sets. Let $K_{t,i,d}$ denote the expected number of distinct $K_{i,j}$'s which are subgraphs of G_t.

Theorem 6. *In the linear growth copying model with constant out-degree d, for $i \leq \log t$,*

$$K_{t,i,d} = \Omega(t \exp(-i)).$$

A new copying model $G(p, \rho, H)$ was recently introduced in Bonato, Janssen [16], motivated by the copying model, the generalized copying graphs of Adler, Mitzenmacher [1], and partial duplication model for biological networks in Chung et al. [23]. The three parameters of the model $G(p, \rho, H)$ are $p \in (0, 1)$, a monotone increasing *random link function* $\rho : \mathbb{N} \to \mathbb{N}$, and a fixed finite initial graph H. The new node v_{t+1} acquires its neighbours as follows. Choose an existing node u from G_t u.a.r.. For each neighbour w of u, independently add an edge from v_{t+1} to w with probability p. In addition, choose $\rho(t)$-many nodes from $V(G_t)$ u.a.r., and add edges from v_{t+1} to each of these nodes.

The existing research on models of W deals almost exclusively with finite graphs. However, in the natural sciences, models are often studied by taking the infinite limit. Limiting behaviour can clarify the similarities and differences between models, and show the consequences of the choices made in the model.

Limit behaviour of a deterministic copying model was investigated in [15], and limit behaviour of the $G(p, \rho, H)$ model was studied in [16]. For a positive integer n, a graph is *n-existentially closed* or *n-e.c.* if for each pair of disjoint subsets X and Y of nodes of G with $|X \cup Y| = n$, there exists a node $z \notin X \cup Y$ joined to every node of X and to no node of Y. A graph is *e.c.* if it is n-e.c. for all positive integers n. By a back-and-forth argument, any two countable e.c. graphs are isomorphic. The unique isomorphism type of countable e.c. graphs is the *infinite random graph* R. The graph R takes its

name from the fact that for any fixed $p \in (0,1)$, with probability 1, a graph $G \in G(\mathbb{N}, p)$ is e.c. The graph R has a rich structure, which the interested reader may read more about in the surveys of P. Cameron [20, 21].

If $(G_t : t \in \mathbb{N})$ is a sequence of graphs with G_t an induced subgraph of G_{t+1}, then define the *limit* of the G_t, written

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

by

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

The following result is essentially stated in [16].

Theorem 7. *Fix $p \in (0,1)$, H, and $\rho = \lfloor \alpha t^s \rfloor$, where α and s are non-negative real numbers with $\alpha, s \in [0,1]$ and $\alpha + p < 1$. Let $G = \lim_{t \to \infty} G_t$ be generated according to the model $G(p, \rho, H)$.*

1. *If $s = 1$ and $\lfloor \alpha t^s \rfloor \geq 1$ for all $t > 0$, then with probability 1 G is isomorphic to R.*
2. *If $s \in [0,1)$ and $\lfloor \alpha t^s \rfloor \geq 1$ for all $t > 0$, then with probability 1 G is $\lfloor \frac{1}{1-s} \rfloor$-e.c.*
3. *If $s \in [0,1)$, then with positive probability G is not isomorphic to R.*

Theorem 7 presents an example of threshold behaviour for convergence to R: with high probability, as s tends to 1, the limit G acquires more and more properties of R, but with positive probability is not itself isomorphic to R. At $s = 1$, we obtain R with high probability.

A new deterministic model for the web graph of note is the *Heuristically Optimized Trade-offs* or *HOT* model of Fabrikant et al. [36]. In the HOT model, nodes correspond to points in Euclidean space, and each node u will link to the node v that performs best in terms of an optimization function which is a linear combination of proximity between u and v, and centrality of v in the network. This implicitly suggests some degree of copying behaviour: a new node whose position is very close to that of an existing node, will have a similar optimization function and hence, is likely to connect to the same node.

7 Growth-deletion models

In all of the models we presented in Sections 4 and 6, at each time step nodes and edges are added, but never deleted. An evolving graph model incorporating in its design both the addition and deletion of nodes and edges may more accurately model the evolution of the web graph. One approach to this was adopted by Bollobás, Riordan [11], who consider the effect of deleting a set of nodes *after* nodes have been generated in the LCD model. The purpose of this study was to investigate the robustness of graphs generated by the LCD model to random failures, and the vulnerability of these graphs to random attack. We now describe two recent models, developed independently of each other, that incorporate the addition and deletion of nodes *during* the generation of nodes. We refer to such models as *growth-deletion models*.

We first describe the growth-deletion model of Chung, Lu [27]. They introduce a model $G(p_1, p_2, p_3, p_4, m)$, with parameters m a positive integer, and probabilities p_1, p_2, p_3, p_4 satisfying $p_1 + p_2 + p_3 + p_4 = 1$, $p_3 < p_1$, and $p_4 < p_2$; the graph H is a fixed nonempty graph. To form $G_{t+1,}$ we proceed as follows. With probability p_1, add v_{t+1} and m edges from v_{t+1} to existing nodes chosen by preferential attachment. With probability p_2, add m new edges with endpoints to be chosen among existing nodes by preferential attachment. With probability p_3, delete a node chosen u.a.r. With probability p_4, delete m edges chosen u.a.r.

By coupling with off-line random graphs, Chung, Lu [27] prove the following result.

Theorem 8. *1. With probability* 1 *as* $t \to \infty$, *the degree distribution of a graph* G_t *generated by* $G(p_1, p_2, p_3, p_4, m)$ *follows a power law distribution with exponent*

$$\beta = 2 + \frac{p_1 + p_2}{p1 + 2p_2 - p_3 - 2p_4}.$$

2. Suppose $m > \log^{2+\epsilon} n$. *For* $p_2 < p_3 + 2p_4$, *we have* $2 < \beta < 3$. *With probability* 1 *as* $t \to \infty$, G_t *is connected with*

$$diam(G_t) = \Theta(\log t)$$

and

$$L(G_t) = O\left(\frac{\log\log t}{\log(1/(\beta-2))}\right).$$

3. *Suppose* $m > \log^{2+\epsilon} n$. *For* $p_2 \geq p_3 + 2p_4$, *we have* $\beta > 3$. *With probability* 1 *as* $t \to \infty$, G_t *is connected with*

$$diam(G_t) = \Theta(\log t)$$

and

$$L(G_t) = O\left(\frac{\log t}{\log d}\right),$$

where d *is the average degree of* G_t.

Another recent growth-deletion model developed independently of [27] is the one of Cooper et al. [29]. The parameters for this model are fixed p_1 and p_2 in $(0,1)$ satisfying $p_2 \leq p_1$, and H is K_1. With probability $1-p_1$ delete a node of G_{t-1} chosen u.a.r. If G_{t-1} has no nodes, then do nothing. With probability p_2, add m edges from v_{t+1} joined to existing nodes chosen by preferential attachment. The graph is made simple by deleting multiple edges. If there are no edges nor nodes in G_{t-1}, then begin again at time $t=0$. If there are no edges but some nodes in G_{t-1}, then add v_{t+1} joined to an existing node chosen u.a.r. With probability $p_1 - p_2$, add m edges between existing nodes, with endpoints chosen by preferential attachment. The graph is made simple by deleting multiple edges and deleting any loops. If there are no edges nor nodes in G_{t-1}, then begin again at time $t=0$. If there are no edges but some nodes in G_{t-1}, then do nothing.

Let $D_k(t)$ be the number of nodes of degree $k \geq 0$ in G_t, and let $\mathbb{E}(D_k(t))$ be the expectation of this random variable. Let

$$\gamma = \frac{2p_1}{3p_1 - 1 - p_2} \text{ and } \rho = \frac{p_2}{p_1}.$$

Cooper et al. [29] prove the following.

Theorem 9. *Assume that* $p_1 + p_2 > 1$. *Then there exists a constant* $C = C(m, p_1, p_2)$ *such that for* $k \geq 1$ *and* $1/2 < p_1 \leq 1$,

$$\left|\frac{\mathbb{E}(D_k(t))}{t} - Ck^{-1-\gamma}\right| = O\left(t^{-\rho/8}\right) + O(k^{-2-\gamma}).$$

As noted in [29], with a suitable choice of p_1 and p_2, we find that γ may take any value in the interval $(1, \infty)$, and so there is a power law for this model with exponent $\beta = 1 + \gamma \in (2, \infty)$.

An intriguing problem is to rigorously analyze the degree distributions of growth deletion models where the choice of nodes and edges to delete is not made u.a.r. A recent model of Flaxman et al. [37] considers an *adversarial* growth deletion model, and analyzes the size of the connected components of graphs generated by the model.

References

1. A. Adler, M. Mitzenmacher, Towards compressing web graphs, In: *Proceedings of the Data Compression Conference*, 2001.
2. W. Aiello, F. Chung, L. Lu, A random graph model for massive graphs, *Experimental Mathematics* **10** (2001) 53-66.
3. W. Aiello, F. Chung, L. Lu, Random evolution in massive graphs, Handbook on Massive Data Sets, (Eds. James Abello et al.), Kluwer Academic Publishers, (2002), 97-122.
4. R. Albert, A. Barabási, H. Jeong, Diameter of the World-Wide Web, *Nature* **401** (1999) 130.
5. R. Albert, A. Barabási, Emergence of scaling in random networks, *Science* **286** (1999) 509-512.
6. P. Baldi, P. Frasconi, P. Smyth, *Modeling the Internet and the Web, Probabilistic Methods and Algorithms*, John Wiley & Sons, Ltd, Chichester, West Sussex, England, 2003.
7. A. Barabási, *Linked: How Everything Is Connected to Everything Else and What It Means*, Perseus Publishing, Cambridge MA, 2002.
8. B. Bollobás, *Modern Graph Theory*, Springer-Verlag, New York 1998.
9. B. Bollobás, *Random graphs, Second edition,* Cambridge Studies in Advanced Mathematics **73** Cambridge University Press, Cambridge, 2001.
10. B. Bollobás, O. Riordan, Mathematical results on scale-free graphs, Handbook of graphs and networks, S. Bornholdt, H. Schuster (eds) Wiley-VCH, Berlin (2002).
11. B. Bollobás, O. Riordan, Robustness and vulnerability of scale-free random graphs, *Internet Mathematics* **1** (2003) 1-35.
12. B. Bollobás, O. Riordan, The diameter of a scale-free random graph, *Combinatorica* **24** (2004) 5-34.
13. B. Bollobás, C. Borgs, T. Chayes, O. Riordan, Directed scale-free graphs, submitted.
14. B. Bollobás, O. Riordan, J. Spencer, G. Tusnády, The degree sequence of a scale-free random graph process, *Random Structures Algorithms* **18** (2001) 279-290.
15. A. Bonato, J. Janssen, Infinite limits of copying models of the web graph, *Internet Mathematics* **1** (2003) 193-213.
16. A. Bonato, J. Janssen, Limits and power laws of models for the web graph and other networked information spaces, accepted to the *Proceedings of Combinatorial and Algorithmic Aspects of Networking*, 2004.

17. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, *Computer Networks* **33** (2000) 309-320.
18. P.G. Buckley, D. Osthus, Popularity based random graph models leading to a scale-free degree sequence, submitted.
19. G. Caldarelli, P. De Los Rios, L. Laura, S. Leonardi, S. Millozzi, A study of stochastic models for the Web Graph, Technical Report 04-03, dipartimento di Informatica e Sistemistica, Universita' di Roma "La Sapienza", 2003.
20. 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.
21. P.J. Cameron, *The random graph revisited*, in: European Congress of Mathematics, Vol. I (Barcelona, 2000), 267–274, Progr. Math., 201, Birkhäuser, Basel, 2001.
22. S. Chakrabarti, *Mining the Web, Discovering Knowledge from Hypertext Data*, Morgan Kauffman Publishers, San Francisco, 2003.
23. F. Chung, G. Dewey, D.J. Galas, L. Lu, Duplication models for biological networks, *Journal of Computational Biology* **10** (2003) 677-688.
24. F. Chung, L. Lu, Connected components in random graphs with given degree sequences, *Annals of Combinatorics* **6** (2002) 125-145.
25. F. Chung, L. Lu, The average distances in random graphs with given expected degrees, *Internet Mathematics* **1** (2003) 91-114.
26. F. Chung, L. Lu, The small world phenomenon in hybrid power law graphs, *Complex Networks*, (Eds. E. Ben-Naim et. al.), Springer-Verlag (2004) 91-106.
27. F. Chung, L. Lu, Coupling on-line and on-line analyses for random power law graphs, submitted.
28. C. Cooper, A. Frieze, On a general model of web graphs, *Random Structures Algorithms* **22** (2003) 311–335.
29. C. Cooper, A. Frieze, J. Vera, Random deletions in a scale free random graph process, submitted.
30. R. Diestel, *Graph theory*, Springer-Verlag, New York, 2000.
31. S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of networks: From biological nets to the Internet and WWW*, Oxford University Press, Oxford, 2003.
32. S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Structure of growing networks with preferential linking, Physical Review Letters **85** (2000) 4633-4636.
33. E. Drinea, M. Enachescu, M. Mitzenmacher, Variations on random graph models for the web, technical report, Department of Computer Science, Harvard University, 2001.
34. P. Erdös, A. Rényi, On random graphs I, *Publ. Math. Debrecen* **6** (1959) 290–297.
35. P. Erdös, A. Rényi, On the evolution of random graphs, *Publ. Math. Inst. Hungar. Acad. Sci.* **5** (1960) 17–61.
36. A. Fabrikant, E. Koutsoupias, 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.
37. A. Flaxman, A. Frieze, J. Vera, Adversarial deletions in a scale free random graph process, submitted.
38. E.N. Gilbert, Random graphs, *Annals of Mathematical Statistics* **30** (1959) 1141-1144.
39. M.R. Henzinger, Algorithmic challenges in web search engines, *Internet Mathematics* **1** (2003) 115-126.
40. S. Janson, T. Luczak, A. Ruciński, *Random Graphs*, John Wiley and Sons, New York, 2000.

41. J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, The web as a graph: Measurements, models and methods, In: *Proceedings of the International Conference on Combinatorics and Computing*, **1627** in LNCS, Springer-Verlag, 1999.
42. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Trawling the web for emerging cyber-communities, In: *Proceedings of the 8th WWW Conference*, 1999.
43. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal, Stochastic models for the web graph, In: *Proceedings of the 41th IEEE Symp. on Foundations of Computer Science*, 57-65, 2000.
44. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal, The web as a graph, In: *Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems*, Publ., Dordrecht, 2002.
45. M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, *Internet Mathematics* **1** (2003) 226-251.
46. S.H. Strogatz, D.J. Watts, Collective dynamics of 'small-world' networks, *Nature* **393** (1998) 440-442.