Stochastic models for the web graph

Ravi Kumar* Prabhakar Raghavan† Sridhar Rajagopalan* D Sivakumar*
Andrew Tomkins* Eli Upfal‡

Abstract

The web may be viewed as a directed graph each of whose vertices is a static HTML web page, and each of whose edges corresponds to a hyperlink from one web page to another. In this paper we propose and analyze random graph models inspired by a series of empirical observations on the web.

Our graph models differ from the traditional $G_{n,p}$ models in two ways:

1. *Independently chosen edges do not result in the statistics (degree distributions, clique multitudes) observed on the web. Thus, edges in our model are statistically dependent on each other.*

2. *Our model introduces new vertices in the graph as time evolves. This captures the fact that the web is changing with time.*

Our results are two fold: we show that graphs generated using our model exhibit the statistics observed on the web graph, and additionally, that natural graph models proposed earlier do not exhibit them. This remains true even when these earlier models are generalized to account for the arrival of vertices over time. In particular, the sparse random graphs in our models exhibit properties that do not arise in far denser random graphs generated by Erdös-Rényi models.

1. Introduction

The web may be viewed as a directed graph in which each vertex is a static HTML web page, and each edge is a hyperlink from one web page to another. Current estimates suggest that this graph has roughly a billion vertices, and an average degree of about 7. In this paper we propose and analyze a class of random graph models inspired by a series of empirical observations on the web graph [5, 11]. These observations suggest that the web is not well modeled by traditional random graph models such as $G_{n,p}$. For instance, the distributions of in- and out-degrees on the web follow a power-law (rather than a Poisson or binomial distribution, as one might expect of a random sparse graph chosen from $G_{n,p}$). Further, it is known [11] that there are several hundred thousand disjoint instances of bipartite cliques ($K_{i,j}$ with $i, j \geq 3$) on the web—once again, an unlikely occurrence in a traditional sparse random graph. Finally, the web is an *evolving* graph: new vertices and edges appear over time, while some older vertices and edges disappear.

We propose a family of random graph models here, very different from the traditional Erdös-Rényi random graph model and its derivatives. Two salient features of our models are worth highlighting here: (1) Because independently chosen edges out of each vertex will not result in the statistics (degree distributions, clique multitude) observed on the web, our model must allow dependencies between edge choices. We achieve this in a simple and plausible manner: some vertices choose their outgoing edges independently at random, as in $G_{n,p}$, but other vertices replicate existing linkage patterns by "copying" edges from a randomly-chosen vertex. We will discuss this further in Section 2. (2) Our model introduces new vertices in the graph as time evolves, to capture the fact that the web is a changing and growing graph. To our knowledge, the only prior work studying the evolution of vertices in the traditional $G_{n,p}$ setting is [2], where the focus is on the emergence of the giant component.

We show that a graph model with the above two features predicts certain graph properties observed on the web. There is an obvious "evolving"[1] version of $G_{n,p}$. Indeed, might it not be possible that such an evolving $G_{n,p}$ (with-

*IBM Almaden Research Center, 650 Harry Road, San Jose CA 95120.
†Verity Inc., 892 Ross Drive, Sunnyvale, CA 94089. Portions of this work were done while the author was at IBM's Almaden Research Center.
‡Computer Science Department, Brown University, Providence, RI 02906. This work was supported in part by the Air Force and the Defense Advanced Research Projects Agency of the Department of Defense under grant No. No. F30602-00-2-0599, and by an NSF grant CCR-9731477.

[1] In this paper, "evolution" in our random graph models refers to the evolution of the graph on the time-axis, rather than on the axis of edge density, as in the seminal work of Erdös and Rényi. This clash of terminology is unfortunate, but the word evolution describes our setting accurately.

out dependencies between the edges) could give rise to the statistical phenomena observed in the web? We show that this is not the case: while an evolving $G_{n,p}$ model behaves very differently from the traditional $G_{n,p}$, the difference is not acute enough to give rise to some of the phenomena observed on the web.

Related work. Kumar *et. al.* [11] describe methods for enumerating subgraphs of the web in the context of discovering web communities. From a graph-theoretic standpoint, a central finding in this work is the existence of a surprising number of edge-induced complete bipartite graphs in the web. The authors also observe the *power-law* distribution of in- and out-degrees in the web graph: the probability that the in-degree of a random vertex is i is distributed by the power-law, $\Pr_u[\text{in-degree}(u) = i] \sim 1/i^\beta$, for $\beta \approx 2.1$. These observations are based on a web crawl from 1997. Other authors [1, 5] verify these degree distributions in more recent web crawls. Interestingly, the power-law exponent in the later experiments is the same as that from the earlier work, suggesting that it may be a fairly stable property of the web graph.

Perhaps the first rigorous effort to define and analyze a model for power-law distributions is due to Herbert Simon [15]. Power-law distributions have been observed for citations in the academic literature, an observation originally due to Lotka [14]. Gilbert [9] presents a probabilistic model supporting Lotka's law. His model is similar in spirit to ours, though different in details and application. The field of bibliometrics [6, 8] is concerned with citation analysis; some of these insights have been applied to the web as well [13].

The "copying" models analyzed in this paper were first introduced by Kleinberg *et. al.*[10]. Motivated by observations of power-laws for degrees on the graph of telephone calls, Aiello, Chung, and Lu [3] propose a model for "massive graphs" (henceforth the "ACL model"), which is very different from ours in three key respects:

- The ACL model ensures the power-law for degrees by first fixing the degrees of (the appropriate number of) vertices to fit the distribution, then randomly introducing edges into the resulting "ports" at each vertex. Thus, the power-law for degrees is an intrinsic feature of the model, rather than an emergent feature of a stochastic process.

- The ACL model was developed to capture characteristics of large-scale call graphs, while ours was developed to capture the nature of the web; thus, their models do not explain the abundance of bipartite cliques observed in the web graph, whereas ours do. See Section 4 for details.

- With vertex degrees being prescribed before any edges are introduced, it is not clear how their model should be adapted to capture the notion of an evolving graph.

Motivations for modeling the web graph.

1. Many problems we wish to solve on the web (such as the subgraph enumeration problems of [12]) are computationally difficult for general graphs. Nevertheless, a suitable model of the web can help us design and analyze algorithms that work well in practice. They could also be simulated under the model to determine their scalability and performance.

2. The model can suggest unexpected properties of today's web that we can then verify and exploit.

Results and organization. In Section 2, we propose our new models that incorporate evolving graphs in which edges are introduced by stochastic copying. We study two variants of these evolving copying models: *linear growth*, in which the graph grows by some absolute amount (i.e., one vertex) at each timestep, and *exponential growth*, in which the graph grows by an amount that depends on its current size (e.g., twice) at each timestep. We also introduce the *evolving uniform* model, in which the graph evolves over time, but edge destinations are chosen independently at random (loosely referred to above as "evolving $G_{n,p}$").

In Section 3 we study the degree distributions in each of these models. Whereas the copying-based models gives rise to power-law distributions, we show that the evolving uniform model has a much flatter degree distribution.

Next, in Section 4, we study the number of bipartite cliques in each of these models, as well as in the ACL model [3]. Bipartite cliques are an interesting class of subgraphs on the web since they capture the notion of "communities" [11]. We show that whereas evolving copying models give rise to large numbers of bipartite cliques (as observed in the web graph), the number of such cliques in the evolving uniform and ACL models is likely to be small. We conclude (Section 5) with a number of directions for further work on modeling and analyzing evolving graphs with and without copying.

2. Random graph models

In this section we give terminology and describe the random graph models we will study. Let $G = \langle V, E\rangle$ denote a directed graph with vertex set V and edge set E. For a directed edge (u, v), u is called the *tail* and v the *head* of the edge. For a vertex u, the edges for which u is the tail (head) are called out-links (in-links) of u. In-degree and out-degree of a vertex are denoted I_u and O_u respectively. The degree of a vertex u in an undirected graph is denoted d_u.

In all our models, we assume the average vertex degree is a constant. This is in light of our focus on the web graph, where we find that despite small average degree, one encounters structures that only arise in far denser graphs in the Erdös-Rényi style of random graphs. For a finite set X, let $x \in_R X$ denote a uniform random choice from X, and for a distribution D let $x \sim D$ denote that x is chosen from the distribution D. Let $[n] = \{1, \ldots, n\}$.

2.1. Evolving graph models

In all of our evolving graph models, the directed graph evolves over discrete timesteps $t = 1, 2, \ldots$. Let the vertices be numbered $1, 2, \ldots$, and let the graph at time t be $G_t = \langle V_t, E_t \rangle$. Two functions are required to describe the evolution of the graph in a model. The growth of vertices is captured by a (possibly random) function $f_v(V_t, t)$ which returns an integer denoting the number of vertices to be added at time $t + 1$; therefore $|V_{t+1}| = |V_t| + f_v(V_t, t)$. The growth of edges is more complicated and is described by a probabilistic edge process $f_e(f_v, G_t, t)$. This function returns the set of edges to be added at time $t + 1$; therefore, $E_{t+1} = E_t \cup f_e(f_v, G_t, t)$. An evolving graph model is completely characterized by $\langle f_v, f_e \rangle$.

Evolving copying models. We consider two different models—*linear growth copying* and *exponential growth copying* models. We begin with an intuitive description of the two models in the context of the web. On the web, pages arrive over time, and page creators link to existing content. We must determine which existing content page creators will have access to in their decisions about which hyperlinks to add. If we assume that web pages are immediately available at creation to the entire browsing population then a page creator should be able to add an edge to any prior vertex. This is *linear growth*: at timestep t, a single vertex arrives and may link to any of the first $t - 1$ vertices. It is reasonable however to assume that a page creator may not be aware of pages created in the last week or two (say). Since the web is currently growing exponentially, this means that a page creator will not see the most recent "epoch" of pages. This is *exponential growth*: at timestep t a new epoch of vertices arrives whose size is a constant fraction of the current graph. Each of these vertices may link only to vertices from previous epochs. We now present the formal definitions.

The linear growth copying model is parameterized by a *copy factor* $\alpha \in (0, 1)$ and a constant out-degree $d \geq 1$. At each time step, one vertex u is added, so $f_v(V_t, t) = 1$, and u is then given d out-links for some constant d. To generate the out-links, we begin by choosing a "prototype" vertex $p \in_R V_t$. The i-th out-link of u is then chosen as follows. With probability α, the destination is chosen uniformly at random from V_t, and with the remaining probability the out-link is taken to be the i-th out-link of p. Thus, the prototype is chosen once in advance. The d out-links are chosen by α-biased independent coin flips, either randomly from V_t, or by copying the corresponding out-link of the prototype.

The intuition behind this model is the following. When an author decides to create a new web page, the author is likely to have some topic in mind. The choice of prototype represents the choice of topic—larger topics are more likely to be chosen. The Bernoulli copying events reflect the following intuition: a new viewpoint about the topic will probably link to many pages "within" the topic (i.e., pages already linked-to by existing resource lists about the topic), but will also probably introduce a new spin on the topic, linking to some new pages whose connection to the topic was previously unrecognized.

The exponential growth model is parameterized by a constant growth factor $p > 0$, the "self-loop" (integral) factor $\gamma > 1$, the "tail copy" factor $\gamma' \in (0, 1)$, and out-degree factor $d > 0$. In this model, degree sequences evolve as a branching process. Let $f_v(V_t, t) \sim B(V_t, p)$, the standard binomial distribution. This branching process has a nonzero extinction probability, but conditioning the process on the fact that it did not terminate, for large t, V_t is well concentrated around its mean, $(1 + p)^t$. To simplify the analysis we assume below (deterministically) that $V_1 = 1$ and $V_t = (1 + p)^t$. The expected number of edges generated in time $t+1$ is $(d+\gamma)pV_t$. Each new vertex is generated with γ self-loop edges. The heads and tails of the remaining edges are chosen according to the following process. Let $u \in V_t$. For each edge directed to u at time t, we generate with probability $dp/(d + \gamma)$ a new edge directed to u. Assuming that the expected number of edges at time t is $(d + \gamma)V_t$, the expected number of edges generated in this process is dpV_t. The tails of the new edges generated in this step are distributed as follows: (1) with probability $1 - \gamma'$ a tail of a new edges is chosen uniformly at random from among the pV_t new vertices of this step and (2) with probability γ' the tail of the edge is chosen at random among the vertices created in previous steps, with the vertices chosen with probabilities proportional to their current out-degree. Therefore, together with the new self-loop edges the expected number of edges at time $t + 1$ is $(d + \gamma)V_{t+1}$.

Linear growth variants. For purposes of comparison, we also introduce a linear growth analog of the standard $G_{n,p}$ random graph model. Again, $f_v(V_t, t) = 1$, and the vertex generated at time t has d out-links. The destination of each out-link is chosen uniformly from the existing vertices. In other words, f_e contains d out-links of the form $(t+1, x)$

for $x \in_R V_t$.

2.2. Static models

For purposes of illustration, we describe some static models. All the graphs in this section are undirected.

Uniform random graphs. The most prevalent and well-studied static random graph model is $G_{n,p}$, in which $V = [n]$ and each possible edge (i, j) is present with probability p. See, for instance, [4].

The ACL model. Generally, given a fixed degree sequence, a family of random graph can be defined by choosing uniformly from all graphs with that degree sequence. Aiello, Chung and Lu [3] describe "power-law random graphs" in which the degree sequence is given by a power-law. The distribution of such graphs can be well-approximated constructively as follows: first a degree sequence is obtained, which fixes the number of vertices and edges. Second, a set is constructed with as many copies of each vertex as its degree. Third, a random matching in this set is chosen. And finally, each edge in the matching between a copy of u and a copy of v is added to the original graph as an edge (u, v).

2.3. Extensions to the models

Our evolving models are by no means complete. They can be extended in several ways. First of all, the tails in our models were either static, chosen uniformly from the new vertices, or chosen from the existing vertices proportional to their out-degrees. This process could be made more sophisticated to account for the observed deviations of the out-degree distribution from the power-law distribution [5]. Similarly, the models can be extended to include *death processes*, which cause vertices and edges to disappear as time evolves. A number of other extensions are possible, but we seek to determine the properties of this simple model, in order to understand which extensions are necessary to capture the complexity of the web.

3. Degree distributions

Let $N_{t,k}$ denote the number of vertices u such that $I_u(t) = k$. In this section we obtain the in-degree distributions in various graph models. The expected in-degree distributions in the case of evolving models follow a power-law—the probability that a random vertex has in-degree i is roughly $\text{poly}^{-1}(i)$. Specifically, in the linear case, we show that $E[N_{t,k}] = tk^{-(2-\alpha)/(1-\alpha)}$ and after T steps, $N_{t,k}$ is sharply concentrated about its mean for t up to about $\ln T$. In the exponential case, we show concentration about the mean $E[N_{t,k}] = O(tk^{\log_\mu(1+p)})$ for $t \leq T^{O(1)}$ and $\mu = 1 + pd/(d + \gamma)$. In contrast, for the evolving uniform model, we show $E[N_{t,k}] = O(t \exp(-k/d))$, i.e., exponentially small tails.

3.1. Evolving copying model: The linear case

For simplicity of exposition, we present the case $d = 1$. Note that this is without any loss in generality, since the linear growth process where out-degree $= d$ can be factored into two probabilistic processes—one for *which* vertex a new vertex decides to copy from, and one for *how many* links it copies from that vertex. The first choice (namely, which vertex to copy from) induces a graph that has the same distribution as a graph in the linear growth model with $d = 1$. This is important for clique analyses.

We first present the analysis for $i = 0$, and build upon it to derive the distributions of $N_{t,i}$ for $i > 0$. Our approach is to study the sequence of random variables $E[N_{t,0} \mid N_{t-k,0}]$ for $0 \leq t - k \leq t$, which forms a martingale. Clearly, $E[N_{t,0}] = E[E[N_{t,0}]] = E[N_{t,0} \mid N_{1,0}]$. The random variable $N_{t,0}$ has the following distribution, which follows directly from the linear growth model:

$$N_{t,0} = \begin{cases} N_{t-1,0} & \text{w.p. } \alpha N_{t-1,0}/(t-1) \\ N_{t-1,0} + 1 & \text{w.p. } 1 - \alpha N_{t-1,0}/(t-1) \end{cases}$$

Lemma 1 *Let* $S_{0,0} = 1$, *and for* $k > 0$, *let* $S_{k,0} = S_{k-1,0}(1 - \alpha/(t-k))$. *Then for every* $t \geq 1$ *and* $0 \leq k \leq t$,

$$E[N_{t,0} \mid N_{t-k,0}] = N_{t-k,0} S_{k,0} + \sum_{j=0}^{k-1} S_{j,0}.$$

Proof: Omitted. □

Next, we establish bounded differences for the martingale $E[N_{t,0} \mid N_{t-k,0}]$.

Lemma 2 *For every* $t \geq 1$ *and every* $k < t$,

$$\left| E[N_{t,0} \mid N_{t-k,0}] - E[N_{t,0} \mid N_{t-(k+1),0}] \right| \leq 2.$$

Proof: Omitted. □

Before stating the tail bound by applying Azuma's inequality, we pause to compute the expected value of $N_{t,0}$.

Lemma 3

$$\frac{t+\alpha}{1+\alpha} - \alpha^2 \ln t \leq E[N_{t,0}] \leq \frac{t+\alpha}{1+\alpha}$$

Proof: Note that $E[N_{t,0}] = E[E[N_{t,0} \mid N_{t-k,0}]] = E[N_{t,0} \mid N_{1,0}]$. By Lemma 1, this equals $\sum_{j=0}^{t-1} S_{j,0}$. We bound this sum by first expressing it as the value of a recurrence, which turns out to be easier to bound sharply. Define the quantity $Q_t = 0$, and for $k < t$, let $Q_k = (1-\alpha/k)(1+Q_{k+1})$. By unwinding the two definitions, it is easy to see that $\sum_{j=0}^{t-1} S_{j,0} = 1 + Q_1$. The lemma follows from the following two claims, whose proofs are omitted: (i) (Upper bound for Q_k) For every $k \le t$, $Q_k \le (t-k)/(1+\alpha)$; in particular, $Q_1 \le (t-1)/(1+\alpha)$. and (ii) (Lower bound for Q_k) $Q_1 \ge (t-1)/(1+\alpha) - \alpha^2 \ln t$. □

We summarize the consequences of Lemmas 2 and 3, together with the Azuma inequality in the following theorem.

Theorem 4 *For any $t > 0$,*

$$\frac{t+\alpha}{1+\alpha} - \alpha^2 \ln t \le E[N_{t,0}] \le \frac{t+\alpha}{1+\alpha}$$

and for all $\ell > 0$,

$$\Pr[|N_{t,0} - E[N_{t,0}]| > \ell] < e^{-\ell^2/4t}.$$

Corollary 5 $P_0 \triangleq \lim_{t\to\infty} E[N_{t,0}/t] = 1/(1+\alpha)$.

We now turn to the more general quantity $N_{t,i}$ for $i > 0$. The goal is to show that for a sufficiently large integer T, after T steps, all the quantities $N_{T,0}, N_{T,1}, \ldots, N_{T,i}$ are sharply concentrated about their respective values $P_0, P_1, \ldots, P_i$, for i up to about $\ln T$. The strategy here is as follows: for each t, we will study the martingale $E[N_{t,i} \mid N_{k,i}, N_{*,i-1}]$ for $k < t$ and where $N_{*,i-1}$ is a shorthand for the list $N_{0,i-1}, N_{1,i-1}, \ldots, N_{t,i-1}$. The sequence $E[N_{t,i} \mid N_{0,i}], E[N_{t,i} \mid N_{1,i}], E[N_{t,i} \mid N_{2,i}], \ldots, E[N_{t,i} \mid N_{t-1,i}]$ is not a martingale in itself; however, conditioned on the values for the random variables $N_{0,i-1}, N_{1,i-1}, \ldots, N_{t-1,i-1}$, this sequence forms a martingale, which is our object of study. We first derive an expression for the quantity $E[N_{t,i} \mid N_{1,i}, N_{*,i-1}]$ in terms of the values of the random variables $N_{*,i-1}$. Then we will inductively assume that $N_{s,i-1}/s$ is bounded by $P_{i-1} \pm T^{-a(i-1)}$ for all $s \ge T^{1-b(i-1)}$ and for suitable decreasing functions a and b. The basis for this induction is provided by Theorem 4. Using the inductive assumption, we first show that $(1/t)E[N_{t,i} \mid N_{1,i}, N_{*,i-1}]$ is $P_i \pm T^{-a(i)}$ for all $t \ge T^{1-b(i)}$. Then by applying the Azuma inequality, we prove that all the $N_{t,i}$'s, for $t \ge T^{1-b(i)}$, are sharply concentrated about their mean values with small error probability, thus completing the inductive step. The error probability for each $N_{t,i}$ will be at most $T^{-\ln T}$, so summing over all $t < T$ and all $i < T$ still gives a negligible total error probability.

We begin by stating the stochastic recurrence for $N_{t,i}$ for $i > 0$:

$$N_{t,i} = \begin{cases} N_{t-1,i} - 1 & \text{w.p. } \frac{\alpha N_{t-1,i} + (1-\alpha) i N_{t-1,i}}{t-1} \\ N_{t-1,i} + 1 & \text{w.p. } \frac{\alpha N_{t-1,i-1} + (1-\alpha)(i-1) N_{t-1,i-1}}{t-1} \\ N_{t-1,i} & \text{otherwise} \end{cases}$$

Using techniques similar to the proof of Lemma 1 and Lemma 2 we obtain:

Lemma 6 *For $i \ge 1$ and integers t and $k < t$, define $F_{k,i-1} = N_{t-k,i-1}/(t-k)(\alpha + (1-\alpha)(i-1))$. Let $S_{0,i} = 1$ and for $k \ge 1$, let $S_{k,i} = S_{k-1,i}\left(1 - \frac{\alpha}{t-k} - \frac{(1-\alpha)i}{t-k}\right)$. Then,*

$$E[N_{t,i} \mid N_{t-k,i}, N_{t-k,i-1}, N_{t-(k-1),i-1}, \ldots, N_{t-1,i-1}]$$

$$= N_{t-k,i} S_{k,i} + \sum_{j=0}^{k-1} S_{j,i} F_{j+1,i-1}.$$

Lemma 7 *For $i \ge 1$ and for every $t \ge 1$ and every $k < t$,*

$$\left| E[N_{t,i} \mid N_{t-k,i}, N_{*,i-1}] - E[N_{t,i} \mid N_{t-(k+1),i}, N_{*,i-1}] \right| \le 2.$$

We now proceed to compute the expected values of $N_{t,i}$. While the goal is to give an analogue of Lemma 3, we now need to condition on the event that the random variables $N_{*,i-1}$ take values close to their expectation. As we proceed from $i-1$ to i, we lose a bit both in the accuracy (i.e., the sharpness of the concentration around the mean) and the range of t's for which the concentration holds.

Let $\theta_i \triangleq \alpha + (1-\alpha)(i-1)$ and $\Delta_i \triangleq \alpha + (1-\alpha)i$.

As a first application of the lemma, we compute the limit of $E[N_{t,i}] = E[E[N_{t,i} \mid N_{1,i}, N_{*,i-1}]] = E[E[N_{t,i} \mid N_{*,i-1}]]$ (since $N_{1,i}$ is the fixed value 0). Inductively, we will assume that $\lim_{k\to\infty} E[N_{k,i-1}]/k = P_{i-1}$; the base case is P_0, which, from Corollary 5, equals $1/(1+\alpha)$. Now, $\lim_{k\to\infty} E[E[N_{t,i} \mid N_{*,i-1}]] = \theta_i P_{i-1}(\lim_{k\to\infty} Q_1^1) = \theta_i/(1+\Delta_i)P_{i-1}$. This, and some crude calculations show:

Theorem 8 *For $r > 0$, the limit $P_r \triangleq \lim_{t\to\infty} N_{t,r}/t$ exists, and satisfies*

$$P_r = P_0 \Pi_{i=1}^r \frac{1 + \alpha/(i(1-\alpha))}{1 + 2/(i(1-\alpha))}$$

and

$$P_r = \Theta\left(r^{-\frac{2-\alpha}{1-\alpha}}\right).$$

We finally proceed to show sharp concentration for the values $N_{t,i}$. For convenience of exposition, let $a(i)$ and $b(i)$ be decreasing functions of i such that $b(i) - b(i+1) \ge a(i+1)$ (roughly, $a(i) = b'(i)$); for definiteness, we take $b(i) \approx 1/(\ln i)$ and $a(i) \approx= 1/(i(\ln i)^2)$.

Theorem 9 *For a sufficiently large integer T, after T steps in the linear growth model, with probability at least $1 - T^{-\Omega(\ln T)}$, for every $i > 0$,*

$$P_i - \frac{1}{T^{a(i)}} \le \frac{N_{t,i}}{t} \le P_i + \frac{1}{T^{a(i)}} \quad \text{for every } t > T^{1-b(i)}.$$

In particular (with the choices $b(i) \approx 1/(\ln i)$ and $a(i) \approx 1/(i(\ln i)^2)$), after T steps, with overwhelming probability, $N_{T,i}/T \in [P_i - \delta, P_i + \delta]$ for some small constant $\delta > 0$ and all $i \le \ln T$.

Proof: The proof proceeds in stages. We inductively assume that the statement of the theorem holds for $i-1$, and show that for every $t > T^{1-b(i)}$, the average value of the martingale $E[N_{t,i} \mid N_{*,i}, N_{*,i-1}]$, conditioned on the values of $N_{*,i-1}$ being in the "right range," is bounded by $P_i \pm T^{-a(i)}$. Then, by applying the bounded differences property for these martingales (from Lemma 7), we obtain the sharp concentration result; this implies that for every $t > T^{1-b(i)}$, every one of the values $N_{t,i}$ will be in the "right range," which allows induction to continue.

Thus, let $i > 0$, and assume that the statement of the theorem holds for $i-1$. Now,

$$\begin{aligned} &E[N_{t,i} \mid N_{1,i}, N_{*,i-1}] \\ &\leq \theta_i Q_1^1 (P_{i-1} + T^{-a(i-1)}) + \theta_i (Q_1^1 - Q_1^{T^{1-b(i-1)}}) \\ &\leq \theta_i \left(\frac{t}{1+\Delta_i}\right)(P_{i-1} + T^{-a(i-1)}) + \theta_i \left(\frac{T^{1-b(i-1)}}{1+\Delta_i}\right). \end{aligned}$$

Thus,
$(1/t)E[N_{t,i} \mid N_{1,i}, N_{*,i-1}] \leq P_i + \frac{\theta_i}{1+\Delta_i}(T^{-a(i-1)} + T^{1-b(i-1)}/t)$. It suffices, therefore, to show that the "error term" $(\theta_i/(1+\Delta_i))(T^{-a(i-1)} + (1/t)T^{1-b(i-1)})$ is at most $T^{-a(i)}$ for $t \geq T^{1-b(i)}$. Following a little manipulation (and assuming that $T^{a(i)} = o(T^{a(i-1)})$ and using the fact that $\theta_i/(1+\Delta_i) < 1$), this is equivalent to showing that $T^{-(b(i-1)-b(i))} \leq T^{-a(i)}$, which follows from the definition of a and b. The lower bound on $(1/t)E[N_{t,i} \mid N_{1,i}, N_{*,i-1}]$ is obtained very similarly, and using the same condition on a and b. The first part of the inductive step is now complete, namely we have shown bounds on the expectation of $E[N_{t,i} \mid N_{1,i}, N_{*,i-1}]$ for all suitable t.

By a simple application of Azuma's inequality, using the bounded differences from Lemma 7, we see that the probability that any fixed $N_{t,i}/t$, for $t > T^{1-b(i)}$, deviates from P_i by more than $T^{-a(i)}$ is at most $T^{-\Omega(\ln T)}$. Thus, summing over all $t \leq T$ and $i \leq T$, the error probability is still of the same form. However, when $i \approx \ln T$, the bound $T^{-a(i)}$ becomes a constant (with the choice $a(i) = 1/(i(\ln i)^2)$), and the bounds fail to be interesting. □

3.2. Evolving copying model: The exponential case

We now analyze the degree distribution in the evolving exponential growth copying model. We show,

Theorem 10

$$\frac{D_1(t)}{k^c} \leq E[N_{t,k}] \leq \frac{D_2(t)}{k^c},$$

where $D_1(t)$ and $D_2(t)$ are functions of t, p, γ and d but not k. c is a function of p, γ and d but not of t and k.

Proof: Fix a vertex u and consider $I_u(t)$, the in-degree of u at time t. $I_u(t)$ can be viewed as a branching process that starts with γ vertices and has

$$\mu = 1 + p\frac{d}{d+\gamma} \text{ and } \sigma^2 = p\frac{d}{d+\gamma}(1 - p\frac{d}{d+\gamma}) \leq p\frac{d}{d+\gamma}.$$

Let $\mu^2 > 1+p$.

Then by simple calculations (see, for example, [7]), $E[I_u(t)] = \gamma\mu^t$ and

$$\text{var}[I_u(t)] = \frac{\gamma\sigma^2\mu^{t-1}(\mu^t - 1)}{\mu - 1} = \gamma\mu^{t-1}(\mu^t - 1).$$

Let $\ell = \log_\mu(k/\gamma)$, and let i^* be the minimum integer i such that $(1-\epsilon)\mu^i \geq 1$, for some ϵ such that $\epsilon^2\gamma\mu > 1$.

For $i \geq i^*$

$$\gamma\mu^{\ell+i} - k = \gamma(\mu^{\ell+i} - \mu^\ell) \geq \epsilon\gamma\mu^{\ell+i}.$$

Thus, by Chebyshev inequality,

$$\begin{aligned} &\Pr\left[I_u(\ell+i) < k\right] \\ &\leq \Pr\left[|I_u(\ell+i) - \mu^{\ell+i}| > \mu^{\ell+i} - k\right] \\ &\leq \frac{\gamma\mu^{2(\ell+i)-1} - \gamma\mu^{\ell+i-1}}{\gamma^2(\mu^{\ell+i} - k)^2} \leq \frac{\gamma\mu^{2(\ell+i)-1}}{\epsilon^2\gamma^2\mu^{2(\ell+i)}} = \delta < 1. \end{aligned}$$

$$\begin{aligned} &E[N_{t,k}] \\ &\geq \sum_{j=1}^{t-\ell-i^*} (1-\delta)(1+p)^j \geq (1-\epsilon_1)\frac{(1+p)^{t-\ell-i^*+1}}{2p} \\ &\geq (1-\epsilon_1)\frac{(1+p)^{t-i^*+1}}{2p(1+p)^\ell} \geq \frac{D_1(t)}{(1+p)^\ell} = \frac{D_1(t)}{k^c}, \end{aligned}$$

for $c = \log_\mu(1+p)$ and

$$D_1(t) = \frac{1-\delta}{2p}\gamma^c(1+p)^{t-i^*+1}.$$

Let $j^* = \frac{1}{2}\log_\mu 2$, then for $j \geq j^*$,

$$\begin{aligned} \Pr[I_u(\ell-j) \geq \gamma k] &\leq \frac{\gamma\mu^{2(\ell-j)-1}}{\gamma^2(\mu^\ell - \mu^{\ell-i})} \\ &\leq \frac{1}{\gamma\mu(\mu^{2j}-1)} \leq \frac{2}{\gamma\mu^{2j+1}}. \end{aligned}$$

$$\begin{aligned} &E[N_{t,k}] \\ &\leq \sum_{j=1}^{t-\ell+j^*} (1+p)^j + (1+p)^{t-\ell+j^*} \sum_{j=1}^{\ell-j^*} \frac{2(1+p)^j}{\gamma(\mu^{2j+1})} \\ &\leq \frac{1}{p}((1+p)^{t-\ell+j^*+1} - 1) + (1+p)^{t-\ell+j^*} \sum_{j=1}^{\ell-j^*} \frac{2}{\gamma\mu} \\ &\leq \frac{(1+p)^{t+j^*+1}}{p(1+p)^\ell} + \frac{2t(1+p)^{t+j^*}}{\gamma\mu(1+p)^\ell} \leq \frac{D_2(t)}{k^c}, \end{aligned}$$

where

$$D_2(t) = \frac{\gamma^c}{p}(1+p)^{t+j^*}(1+p+\frac{2pt}{\gamma\mu}),$$

using $\mu^2 > 1+p$. □

This yields the corollary:

Corollary 11 *For t and k such that $\frac{D_1(t)}{k^c}, \frac{D_2(t)}{k^c} \to \infty$, and for any $\epsilon > 0$*

$$\Pr\left[(1-\epsilon)\frac{D_1(t)}{k^c} \le N_{t,k} \le (1+\epsilon)\frac{D_2(t)}{k^c}\right] = 1 - o(1).$$

Proof: The degrees of different vertices are independent random variables. Thus, $N_{t,k}$ is the sum of 0-1 independent random variables. □

3.3. Evolving uniform model

Let $v_1, v_2, \ldots$ be the vertices added at time $t = 1, 2, \ldots$.

Lemma 12 *For $t' < t$, $\mu = E[I_{v_{t'}}(t)] = d\ln(t/t')$ and $\Pr[(1-\delta)\mu \le I_{v_{t'}} \le (1+\delta)\mu] > 1 - 2\exp(-\mu\delta^2/4)$ for sufficiently small $\delta > 0$.*

Proof: The expected increase in in-degree for $v_{t'}$ is given by $\sum_{i=t'}^{t} d/i$, which yields μ. Also, using independence of the choices, the distribution is concentrated around its expectation. □

Corollary 13 $E[N_{t,k}] = O(t\exp(-k/d))$.

Proof: Notice that for all vertices $v_1, \ldots, v_{t\exp(-k/d)}$, the expected degree of each of them is at least k. Hence, $E[N_{t,k}] = t\exp(-k/d) - t\exp(-(k-1)/d = O(t\exp(-k/d))$. The degree distribution is concentrated around the mean since each of vertices has expected degree very close to mean as shown in the previous lemma. □

4. Number of cliques

Recall that $K_{i,j}$ is a bipartite clique when all the ij possible edges are present. Since our random graphs are directed, we consider the situation when the edges are directed from i vertices to j vertices.

In this section, we count the number of bipartite cliques that arise in the different graph models. We also count the number of bipartite cliques in a directed version of the ACL model to show that our evolving copying model is fundamentally different from this model. Let $K(t,i,j)$ denote the expected number of $K_{i,j}$'s present in the graph at time t. In many of the cases, we focus only on $K(t,i,i)$'s. We distinguish the evolving copying models from the other models by showing that in the copying models there are many (t^ϵ) large cliques, while there are only very few cliques in the uniform evolving model, and very few large cliques in the ACL model.

4.1. Evolving copying models

The following theorem shows that there are many cliques in the evolving copying model with linear growth, even with constant out-degree. One can define a variant of the linear growth copying model in which the tails of edges are also chosen by copying processes; for such models, we can show that there are many copies of $K_{i,j}$; we instead focus on $K_{i,d}$'s.

Theorem 14 *In the linear growth copying model with constant out-degree d, for $i \le \log t$, $K_{t,i,d} = \Omega(t\exp(-i))$.*

Proof: Call a vertex v_τ arriving at time $\tau \le t$ a *leader* if at least one of its d out-links is chosen uniformly, i.e., without copying. Notice that a given node is a leader with probability $1-(1-\alpha)^d$. Call a vertex a *duplicator* if it copies all d of its out-links, and note that a node is a duplicator with probability $(1-\alpha)^d$. Now, consider a leader v_τ. Consider the epochs $(\tau, 2\tau], (2\tau, 4\tau], \ldots, (t/2, t]$. The probability that at least one vertex in the first epoch copies from v_τ is at least $1 - \prod_{\tau'=\tau+1}^{2\tau}(1 - 1/(\tau+\tau')) \approx 1/2$, and likewise for subsequent epochs. Thus, the expected number of duplicators of v_τ is $\Omega(\ln(t/\tau))$. The random variable denoting the number of duplicators of v_τ is concentrated about its mean because each epoch is an independent event with constant probability of contributing a duplicator.[2] Now, v_τ and its duplicators form a complete bipartite subgraph.[3] It then follows, for $i \le \log t$, $K_{t,i,d} = \Omega(t\exp(-i))$. □

The following theorem shows that there are a lot of cliques in the evolving copying model, the exponential growth case.

Theorem 15 *There are constants $c = c(p,\mu) < 1$ and $\theta = \theta(p,\mu) < 1$, independent of i and t, such that $K(t+1,i,i) = \Omega((1+p)^{ct}\theta^{i^2})$.*

Proof: (Sketch) We condition on two events that hold with high probability: (1) For some constant $b > 0$ there are at least $b(1+p)^{t-j}$ vertices of degree at least μ^j at time t; (2) For some constant $a > 0$ there are no more than $a(d+\gamma)(1+p)^t$ edges at time t.

Let u and v be two vertices of degree at least μ^j at time t. The probability that a new edge connects u to v at time $t+1$ is at least

$$q = 1 - \left(1 - \frac{\gamma' p d \mu^j}{a(d+\gamma)(1+p)^t}\right)^{\mu^j} = \theta\frac{\mu^{2j}}{(1+p)^t}$$

for some $0 < \theta < 1$.

[2] We can attain better bounds by considering duplicators of duplicators; this formulation yields a branching process similar to the process of Section 3.2.

[3] For $j < d$, we can attain better bounds for $K_{t,i,j}$; for simplicity, we treat d as a constant.

Partition the set of $b(1+p)^{t-j}$ vertices of degree μ^j into $r = (b/i)(1+p)^{t-j}$ disjoint sets of i vertices each. Divide the r sets into two equal size groups V and W. The probability that a given set in V and a given set in W are connected by i^2 edges at time $t+1$, to complete a $K_{i,i}$ is q^{i^2}.

To count disjoint cliques we construct up to $r/2$ cliques; thus each set in V has at least $r/2$ possible sets in W to choose from. Thus, the expected number of disjoint cliques is at least $(r/2)(1-(1-q^{i^2})^{r/2})$. For $j > t(\log(1+p)/(2\log\mu)+o(1))$, $1-(1-q^{i^2})^{r/2} \geq \theta^{i^2}$. □

4.2. Evolving uniform model

Theorem 16 *For* $t > 0, i > e^2+1$, $K(t,i,i) < 2$.

Proof: We assume $i \leq d$. For the formation of a $K_{i,i}$ from vertices $U = \{u_1, \ldots, u_i\}$ to vertices $V = \{v_1, \ldots, v_i\}$, we need all of the i edges emanating from each $u_j \in U$ to link into distinct members of V. For the sake of establishing an upper bound on the expected number of such cliques, we will merely insist that all of the i edges emanating from each u_j link into V, without insisting that they link into distinct members of V. Enumerating over all choices of u_j, v_j, the expected number is bounded from above by

$$\int_{u_1=i}^{\infty} \binom{u_1}{i} \left(\frac{i}{u_1}\right)^i \int_{u_2=u_1}^{\infty} \left(\frac{i}{u_2}\right)^i \cdots \int_{u_i=u_{i-1}}^{\infty} \left(\frac{i}{u_i}\right)^i.$$

The above expression is an upper bound since we omit several $+1$ terms (in the lower limits of the integrals, in the denominators of the probabilities, etc.) and we let the upper limits of the integrals be ∞ rather than t.

We next bound $\binom{u_1}{i}$, the number of ways of choosing V from vertices to the left of u_1 by $(eu_1/i)^i$, and integrate. The expectation is then bounded above by

$$(i^{i-1}e)^i \int_{u_1=i}^{\infty} \int_{u_2=u_1}^{\infty} \left(\frac{1}{u_2}\right)^i \cdots \int_{u_i=u_{i-1}}^{\infty} \left(\frac{1}{u_i}\right)^i$$

Integrating out, the upper bound becomes

$$\begin{aligned} \frac{(i^{i-1}e)^i}{i!(i-1)^{i-1}(i-2)i^{i(i-2)}} &= \frac{(ei)^i}{i!(i-2)(i-1)^{i-1}} \\ &< \frac{e^{2i}}{(i-2)(i-1)^{i-1}}. \end{aligned}$$

In particular, even for $i = 3$, $K(t,i,i) < 23$ and for $i > e^2+1$, this number is under 2. □

4.3. Cliques in the ACL model

Let $K(i,j)$ denote the expected number of $K_{i,j}$'s present in a graph. We compute $K(i,j)$ in a directed version of the ACL model. The ACL model for given $\alpha, \beta > 0$ is the following: assign uniform probability to all graphs with $N(k) = \exp(\alpha)/k^\beta$ (self-loops are allowed), where $N(k)$ is the number of nodes with out-degree k. Let $G = (V,E)$ be generated according to this model. The following lemma can be proved.

Lemma 17 *There is a constant c (slightly above 1) such that* $\Pr_{u,v}[(u,v) \in E] < cd_ud_v/(2E)$.

The following theorem shows that there are very few bipartite cliques in this model.

Theorem 18 *For* $i > 2/(\beta-2)$, $K(i,i)$ *is constant.*

Proof: Computing $K(i,j)$ is equivalent to summing over all i-tuples and j-tuples of vertices, the probability that all the edges exist between them. Let $d_1, d_2, \ldots, d_{i+j}$ be the degrees of vertices. Notice that the maximum degree of a vertex in their model is given by $\exp(\alpha/\beta)$ and the probability that a vertex has degree d is given by $\exp(\alpha)/d^\beta$. Then, the expected value of $K(i,j)$ is upper bounded by the sum

$$\int_{d_1,\ldots,d_{i+j}}^{\exp\frac{\alpha}{\beta}} \left(\prod_{\ell=1}^{i+j} \frac{\exp(\alpha)}{d_\ell^\beta}\right) c^{ij} \left(\prod_{\ell=1}^{i} \frac{d_\ell^j}{2E}\right) \left(\prod_{\ell=1}^{j} \frac{d_\ell^i}{2E}\right).$$

We restrict our attention to $K(i,i)$, in which case the sum is upper bounded by

$$\begin{aligned} &\frac{\exp(2i\alpha)}{(2E)^{i^2}} \int_{d_1,\ldots,d_{2i}}^{\exp\frac{\alpha}{\beta}} (d_1 \ldots d_{2i})^{-(\beta-i)} \\ &= \exp\left((2i^2+2i)\frac{\alpha}{\beta} - i^2\alpha\right). \end{aligned}$$

For $i > 2/(\beta-2)$, this quantity is constant. □

5. Further work

A number of directions for further work arise. (1) Our models allow for the web graph to evolve by the addition of vertices and edges; more generally, we could study models with vertex- and edge-deletion. (2) Some of our evolving models result in directed acyclic graphs; by introducing processes for deleting and re-introducing edges, one can remedy this. What are the effects on the properties of the resulting graphs? (3) Recent heuristic calculations [1] argue that the web graph has a small diameter; on the other hand, observations by Broder *et. al.* [5] suggest that the reality is somewhat more complicated. What light can our models shed on this? (4) What is the size of the connected components of our graph models, and how would this reconcile with the observations of [5]? (5) What can be said of the efficiency of algorithms on evolving and/or copying-based random graphs?

References

[1] R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the World-Wide Web. *Nature* 401:130–131, 1999.

[2] D. Aldous and B. Pittel. On a random graph with immigrating vertices: Emergence of the giant component. *Preprint*, 2000.

[3] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. *Proc. ACM Symp. on Theory of Computing*, pp. 171–180, 2000.

[4] B. Bollobás. *Random Graphs*. Academic Press, 1985.

[5] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. *Proc. 9th WWW Conf.*, pp. 309–320, 2000.

[6] L. Egghe and R. Rousseau. *Introduction to Informetrics*. Elsevier, 1990.

[7] W. Feller. *An Introduction to Probability Theory and its Applications: I*, John Wiley, 1950.

[8] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972.

[9] N. Gilbert. A simulation of the structure of academic science. *Sociological Research Online*, 2(2), 1997.

[10] J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins. The web as a graph: Measurements, models and methods. *Proc. Intl. Conf. on Combinatorics and Computing*, pp. 1–18, 1999.

[11] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Proc. 8th WWW Conf.*, pp. 403–416, 1999.

[12] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. *Proc. VLDB*, pp. 639–650, 1999.

[13] R. Larson. Bibliometrics of the World-Wide Web: An exploratory analysis of the intellectual structure of cyberspace. *Ann. Meeting of the American Soc. Info. Sci.*, 1996.

[14] A. J. Lotka. The frequency distribution of scientific productivity. *J. of the Washington Acad. of Sci.*, 16:317, 1926.

[15] H. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[16] G. K. Zipf. Human behavior and the principle of least effort. *New York: Hafner*, 1949.