The influence of search engines on preferential attachment

Soumen Chakrabarti Alan Frieze* Juan Vera

Carnegie Mellon University
Pittsburgh PA15213

Abstract

There is much current interest in the evolution of social networks, especially, the Web graph, through time. "Preferential attachment" and the "copying model" are well-known models which explain the observed degree distribution of the Web graph reasonably closely. We claim that the presence of highly popular search engines like Google substantially mediate the act of hyperlink creation by limiting the author's attention to a small set of "celebrity" URLs. Page authors (who are also Web surfers) frequently (with probability p) locate pages using a search engine. Then they link to popular pages among those they visit. We initiate an analysis of this more realistic process, and show that the celebrity nodes eventually accumulate a constant fraction of all links created **whp**, and that the degrees of the other nodes still follow a power-law distribution, but with a steeper power: $\mathbf{Pr}(\text{degree} = k) \propto k^{-(1+2/(1-p))}$ **whp**. Our analysis adds evidence to the recent concern that search engines offer new Web pages a steep, self-sustaining barrier to entry to well-connected, entrenched Web communities.

1 Introduction

The evolution of the Web graph through time has been subject to intense modeling, measurements, and analysis in recent years. Early measurements on the graph of Web pages (nodes) and hyperlinks (edges) showed that degrees of nodes were distributed according to a power law. Barabasi and Albert [1] were among the first to propose a generative model of the Web, called *preferential attachment*, which leads to a distribution $\mathbf{Pr}(\text{degree} = k) \propto k^{-3}$.

Kleinberg *et al.* [7] were the first to propose a *copying model* in which the author of a newborn page u picks a random reference page v from the Web, and with some probability, copies out-links from v to u. Kumar *et al.* [8] analyzed the copying process to show that it, too, leads to a power law degree distribution with a power of approximately 2, which is close to empirical observations.

Both these generative models hint that the author of a new page is potentially influenced by all existing pages: she is either influenced by their current degrees, or she can sample a reference page uniformly. Kumar *et al.* also consider a *geometric copying model* in which the Web grows so rapidly that the author of a new page can be influenced only by a fraction of the pages that will have been created by the end of the current time-step. But in absolute terms, this can still translate to billions of pages.

In reality, the evolution of the Web graph has been influenced permanently and pervasively by the existence of search engines. Responses from search engines significantly influence where authors are likely to link. This in turn influences degree and Pagerank, which are used by most search engines to rank their results. Thus, search engines, which started out *observing* social linkage phenomena on the Web, are now *influencing* the outcome.

Consider the uniform "teleport" jump in the well-known *random surfer* model at the heart of Pagerank (which powers Google). According to Neilsen/NetRatings[1], an estimated 319 million searches are answered by 10 major search engines each day. Therefore, it seems more likely that with some significant probability, teleports take the surfer to a search engine (instead of a uniformly random destination), whence the surfer is taken to highly popular pages. Therefore, the teleport has become highly biased, and the original model is in question.

The virtuous cycle of limelight can be brutal to new pages and sites: Cho and Roy [2] estimate that the time taken for a page to reach prominence can be delayed by a factor of over 60 if a search engine diverts clicks to entrenched pages. Drinea *et al.*

*Supported in part by NSF grant CCR-0200945.

[1] http://www.nielsen-netratings.com

[4] analyze balls-and-bins processes with a related feedback mechanism, and show that positive feedback leads to a rapid landslide victory for the winning bin. In a world where copious content jostles for scarce attention, this is not new. Similar effects result from, e.g., the New York Times bestsellers list.

Having some empirical understanding of the effect of search engines on the evolution of page popularity for search applications, we are interested in directly modeling the evolution of the Web graph under the influence of a search engine.

1.1 Our model

We wish to model how the Web graph evolves if authors use search engines to decide on links that they insert in new pages. In particular, we are interested in the degree distribution, and whether and how this distribution deviates from those derived by Barabasi, Kleinberg, Kumar, and co-workers.

For simplicity, like Barabasi *et al.*, we model the Web graph as undirected. Following Cho and Roy, we also make the simplifying assumption that the query to the search engine is fixed and the search engine, like a bestseller list, returns some *fixed number* of response URLs (nodes in the Web graph), ordered according to their degree at the end of the previous time-step. We can also interpret such a list as a per-topic listing provided by a directory like Yahoo! or DMoz, and limit our analysis to one topic at a time, without loss of generality.

The growth process we seek to analyze generates a sequence of graphs $G_t, t = 1, 2, \ldots$. At time t, the graph $G_t = (V_t, E_t)$ has t vertices and mt edges. The process has only two important parameters p (a probability) and N (the maximum number of "celebrity" nodes listed by the search engine).

We introduce some notation:

$deg_t(x)$ denotes the degree of vertex x in G_t

$D_t(U)$ is $\sum_{x \in U} deg_t(x)$

S_t denotes the set of at most N vertices with the largest degrees in G_t. (If $t < N$ we let $S_t = V_t$.)

$d_k(t)$ denotes the number of vertices of degree k at time t in the set $V_t - S_t$.

$\bar{d}_k(t)$ is defined as $\mathbf{E}[d_k(t)]$, the expectation being over the random hyperlinking choices made by nodes (described next)

The graph sequence is constructed as follows:

Time step 1: The process is initialized with graph G_1 which consists of an isolated vertex x_1 and m loops.

Time step $t > 1$: We add a vertex x_t to G_{t-1}. We then add m random edges (x_t, y_i), $i = 1, 2, \ldots, m$ incident with x_t, where y_i are nodes in G_{t-1}. For each i:

- With probability p we choose $y_i \in S_{t-1}$.
- With probability $q = 1 - p$ we choose $y_i \in V_{t-1}$.

In both cases y_i is selected by preferential attachment within the target subset of old nodes, i.e. for $x \in U$

$$\mathbf{Pr}(y_i = x) = \frac{deg_{t-1}(x)}{D_{t-1}(U)},$$

where $U = S_{t-1}$ or $U = V_{t-1}$ as the case may be.

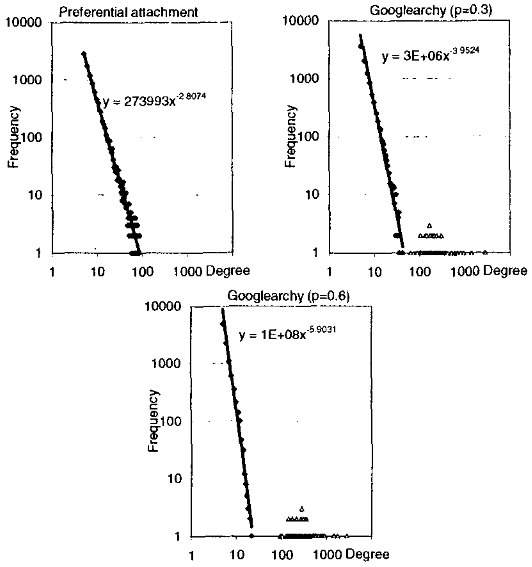
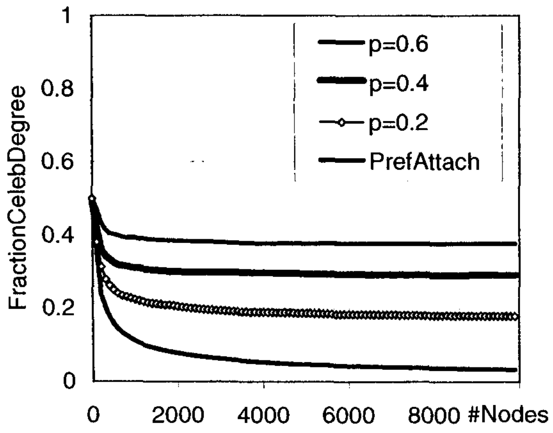


Figure 1: The presence of a search engine in our model makes the power in the degree power law more negative, and, with increasing p, separates out the celebrities completely from the non-celebrities ($N = 100$, $n = 10000$, and $m = 5$).

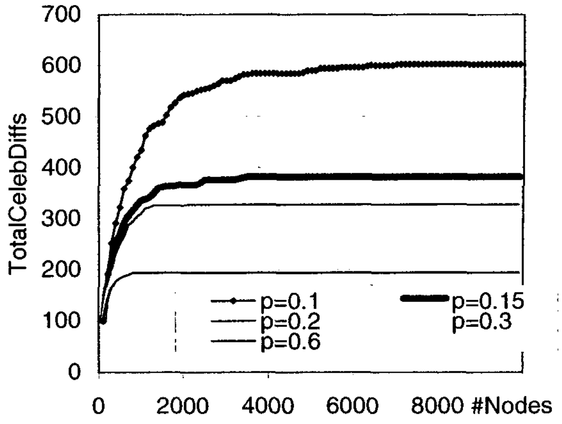
As Figure 1 shows, the simulated behavior of our proposed process is quite different from standard preferential attachment. With increasing p, the celebrities swing out far from the power-law straight line in log-log plots.

Furthermore, as Figure 2 shows, the total degree (as a fraction of twice the total number of edges added) over the celebrities goes to zero as $n \to \infty$ for



$N = 10$, $n = 10000$, and $m = 5$

Figure 2: The total degree of the celebrities as a fraction of (twice) the number of edges added to the graph differs significantly in behavior between preferential attachment vs. our model.



$N = 100$, $n = 10000$, and $m = 5$

Figure 3: The celebrity list becomes effectively fixed very early on in the graph evolution process and the cumulative number of celebrity shuffles levels out faster with large p.

preferential attachment, but in a simulation of our proposed model, the celebrities command a *constant fraction* of the total degree over all nodes, and this fraction grows with p. In Figure 3 we plot the cumulative number of nodes leaving or entering the celebrity list from each timestep to the next. We see that as p increases, the celebrity list is determined more and more quickly.

As we shall see, the observations above lend much intuition to the analysis of our proposed graph evolution process.

1.2 Our results and their implications

We will prove the following, where all asymptotic notation is with respect to n:

THEOREM 1.1.

(a) *For every $i \leq N, \mathbf{E}[\deg_n(x_i)] = \alpha_i n + O(n^{1/2})$ for some constant $\alpha_i > 0$. I.e., each celebrity commands a constant fraction of all edges ever generated in the graph.*

(b) *There is an absolute constant A_1 such that for every $k \geq m, \bar{d}_k(n) = (1 + o(1))\frac{A_1 n}{k^{1+2/q}}$.*

Our analysis involves a coupled sequence of graphs, G_t^*, $t = 1, 2, \ldots$, obtained by the analogous process to the one above, where in each step S_t is replaced by $S_t^* = S^* = \{x_1, \ldots, x_N\}$. (If $t < N$ take $S_t^* = V_t$.) I.e., instead of taking the N largest-degree vertices, we take the N *oldest* vertices.

Our model differs from reality in many obvious ways: edges are undirected, outlinks are not modified after creation, pages do not die, and there is no topic-based clustering. Yet, our results lend support to recent articles by political scientists [6] in the popular press expressing apprehension about the extent to which search engines concentrate the collective attention of Web surfers to "mainstream" Web sites.

2 Coupling G_t and G_t^*

Let m_t be the degree of the lowest degree vertex in S_t and M_t the degree of the highest degree vertex in $V_t \setminus S_t$. We are going to prove that after a short time **whp** there is a significant gap between m_t and M_t and then from this time on S_t the set of the N highest degree vertices remains fixed. In this sense the graph G_t is very similar to the graph G_t^* where the top N is fixed from the beginning (the top is fixed by age not by degree). We define m_t^* and M_t^* for G_t^* in an analogous way to m_t and M_t.

LEMMA 2.1. *Conditional on $S_t = S$ and $D_t(S_t) = D$, the distribution of degrees $V_t \setminus S$ is identical with the distribution of degrees in $V_t \setminus S_t^*$ conditional on $D_t^*(S_t^*) = D$.*

Proof The only difference between the generation of edges in G_t incident with $V_t \setminus S_t$ is that occasionally a vertex x from $V_t \setminus S_t$ replaces a vertex y in S_t. From now on, as far as the degree sequence of $V_t \setminus S_t$ is concerned, this is equivalent to re-labelling

x with y, even though the edge structure will change. □

Lemma 2.2. *We can couple G_t and G_t^* in such a way that $D_t(V_t \setminus S_t) \leq D_t^*(V_t^* \setminus S_t^*)$ and so $M_t^* \geq M_t$ in distribution.*

Proof We construct G_k and G_k^* simultaneously $k = 1, 2, \ldots, t$ with $G_k = G_k^*$ for $k = 1, 2, \ldots, N$. In general, given G_k, G_k^*, we add vertex x_{k+1} to both. We assume that $D_k(S_k) \geq D_k^*(S_k^*)$ and then for $i = 1, 2, \ldots, m$ we choose its neighbours y_i, y_i^* as follows: With probability p we choose y_i preferentially from S_k and y_i^* preferentially from S_k^*. These choices are done independently. With probability $1 - p$ we choose both preferentially from V_k, with the proviso that if $y_i^* \in S_k^*$ then $y_i \in S_k$. Note that sometimes y_i will move into S_k replacing some vertex x. Since y_i, x had the same degree before the addition of an edge, this coupling has the desired properties. □

Lemma 2.3. *We can couple G_t and G_t^* in such a way that $m_t^* \leq m_t$ in distribution.*

Proof For $t \geq N$ the degrees of the vertices in S_t follow an urn model. In each step either (i) we add a ball (endpoint of the edge x_t) and place it in an urn according to urn size or (ii) we add a ball to the smallest urn (a vertex moves into S_t replacing another vertex). If we replace (ii) by simpling adding a ball as in method (i) then we can couple the two processes so that in the former process the smallest urn size is at least the smallest urn size in the latter. The latter process corresponds to G_t^*, but with possibly more balls going into S_t^*. □

Proof of Theorem 1.1
Let ρ be the last time that S_t changes in the G_t process. It follows from Lemma 3.3 (below) that

$$\text{(2.1)} \qquad \mathbf{Pr}(\rho \geq t) \leq \epsilon_t \qquad \text{where } \lim_{t \to \infty} \epsilon_t = 0.$$

From time $t \geq \rho$, S_t is fixed. Condition on $\rho \leq \ln n$ and the degrees $\mathbf{d} = (d_1 \geq d_2 \geq d_N)$ in S_t at this point. The degrees at time n will be identical in distribution to the contents of N urns, with initial contents $\mathbf{d}$ into which $\sim \frac{2mp}{1+p} n$ (see Lemma 3.4) balls have been randomly placed according to a Polya-Eggenburger scheme [9].

As such, the expected degrees of the contents of urn i can be expressed as $\sim \psi_i(\mathbf{d}, m, p)n$. Thus we can prove part (a) of the theorem if we can argue that

$$\alpha_i = \sum_\rho \sum_{\mathbf{d}} \psi_i(\mathbf{d}, m, p) \mathbf{Pr}(\rho, \mathbf{d}) > 0.$$

But $\alpha_N > 0$ follows immediately from (2.1) or from Lemmas 2.3 and 3.2. (Note that α_i will be different from the expresion $\alpha_i^* = \frac{mt}{N} \prod_{1 \leq j < i} \left(1 + \frac{1}{2j}\right)$ given in Lemma 3.2, due to differences in the early growth of G_t, G_t^*. We do know however that $\alpha_N \geq \alpha_N^*$).

Whp the G_n degree distribution of $V_n \setminus S_n$ can be described as follows: Up to time ρ, in distribution, fewer edges are created with endpoints chosen preferentially than in G_n^*. After this time, the remaining edges are created in the same way as in G_t^*. Define the event

$$\mathcal{E} = \bigcap_{t=1}^{n} \{M_t \leq Kt^{q/2}(\ln t)^3)\}$$

where K is some large constant.

The conclusion of Lemma 3.7 is also valid for G_t and so $\mathbf{Pr}(\overline{\mathcal{E}}) = O(t^{-\kappa})$ for any constant $\kappa > 0$. From Lemma 2.1, the two processs coincide from time $\ln n$ onwards **whp** and we can apply Lemma 3.1 since we can assume $\mathcal{E}^*$ holds (equivalent event to $\mathcal{E}$ in the context of G_t^*). □

3 Analysis of G_t^*

In this section we analyze the behavior of G_t^*. In Lemma 3.1 we prove that $\overline{d_k^*}(t)$ follows a power law, while in Lemma 3.2 we prove that $deg_n^*(x_i)$ is linear for $i \leq N$. Then we turn our attention to computing different parameters of G_t^*. Let

$$C_N = \frac{2N^{\frac{1+p}{2}}}{1+p}.$$

Define

$$\mathcal{E}^* = \bigcap_{t=1}^{n} \{M_t^* \leq Kt^{q/2}(\ln t)^3)\}$$

where K is some large constant.

Lemma 3.1. *Let $t_0 = \ln n$, fix $G_{t_0}^*$ and assume $k \geq m$. Condition on $\mathcal{E}^*$. Then*

$$\overline{d_k^*}(n) = (1 + o(1)) \frac{A_1 n}{k^{(1+2/q)}}.$$

Proof Our approach to proving a power law is to find a recurrence for $\overline{d_k^*}(t)$. Lemma 3.7 shows that $\mathbf{Pr}(\overline{\mathcal{E}^*}) = O(t^{-K})$ for any constant $K > 0$.

Thus corrections due to conditioning can easily be absorbed into the error term.

We define $\overline{d^*_{m-1}}(t) = 0$ for all $t > 0$. Then for $t \geq t_0, k \geq m$,

$$\begin{aligned}
&\mathbf{E}\left[d^*_k(t+1) \mid G_t\right] \\
&\quad = d^*_k(t) + qm\left(\frac{(k-1)d^*_{k-1}(t)}{2mt} - \frac{kd^*_k(t)}{2mt}\right) \\
&\qquad + 1_{k=m} + O(M^*_t t^{-1}) \\
&\quad = d^*_k(t) + q\frac{(k-1)d^*_{k-1}(t) - kd^*_k(t)}{2t} \\
&\qquad + 1_{k=m} + O(M^*_t t^{-1}).
\end{aligned}$$

The $O(M^*_t t^{-1})$ term accounts for the addition of parallel edges.

Taking expectations, we get

$$\begin{aligned}
\overline{d^*_k}(t+1) &= \overline{d^*_k}(t) + q\tfrac{(k-1)\overline{d^*_{k-1}}(t) - k\overline{d^*_k}(t)}{2t} \\
&\quad +1_{k=m} + O(t^{q/2-1}(\ln t)^3). \qquad (3.2)
\end{aligned}$$

We consider the exact recurrence, $f_{m-1} = 0$ and

$$f_k = 1_{k=m} + q\frac{(k-1)f_{k-1} - kf_k}{2} \qquad \text{for } k \geq 0, \qquad (3.3)$$

yielding

$$\begin{aligned}
f_k &= f_m \prod_{i=m+1}^{k} \frac{i-1}{i+2/q} \\
&\sim f_m k^{-(1+2/q)}.
\end{aligned}$$

We finish the proof of the lemma by showing that there exists a constant $M > 0$ such that

$$|\overline{d^*_k}(t) - f_k t| \leq M(t_0 + t^{q/2}(\ln t)^3) \qquad (3.4)$$

for all $t > 0$.

Let $\Theta_k(t) = \overline{d^*_k}(t) - f_k t$. Then for $k \geq m$ and $t \geq t_0$,

$$\begin{aligned}
\Theta_k(t+1) &= \left(1 - \tfrac{qk}{2t}\right)\Theta_k(t) + \tfrac{q(k-1)}{2t}\Theta_{k-1}(t) \\
&\quad +O(t^{q/2-1}(\ln t)^3). \qquad (3.5)
\end{aligned}$$

Let L denote the hidden constant in $O(t^{q/2-1}(\ln t)^3)$ of (3.5). Our inductive hypothesis $\mathcal{H}_t$ is that $|\Theta_k(t)| \leq M(t_0 + t^{q/2}(\ln t)^3)$ for every $k \geq m$. It is trivially true for $t \leq t_0$. So assume that $t \geq t_0$. Then, from (3.5),

$$\begin{aligned}
|\Theta_k(t+1)| &\leq M(t_0 + t^{q/2}(\ln t)^3) + Lt^{q/2-1}(\ln t)^3 \\
&\leq M(t_0 + (t+1)^{q/2}(\ln t)^3)
\end{aligned}$$

provided $M \geq 2L$. This verifies $\mathcal{H}_{t+1}$ and completes the proof by induction. □

Lemma 3.2. *For $i \leq N$ and $t \geq N$,*

$$\mathbf{E}\left[deg^*_t(x_i)\right] = \frac{mt}{N}\prod_{1 \leq j < i}\left(1 + \frac{1}{2j}\right) + \tilde{O}(t^{1/2})$$

Proof Let $t \geq N$, then

$$\begin{aligned}
\mathbf{E}\left[deg^*_{t+1}(x_i) | G^*_t\right] &= deg^*_t(x_i) + mp\frac{deg^*_t(x_i)}{D^*_t(S^*_t)} \\
&\quad + mq\frac{deg^*_t(x_i)}{2mt}.
\end{aligned}$$

Taking expectations we get

$$\begin{aligned}
\mathbf{E}\left[deg^*_{t+1}(x_i)\right] &= \mathbf{E}\left[deg^*_t(x_i)\right]\left(1 + \frac{q}{2t}\right) \\
&\quad + mp\mathbf{E}\left[\frac{deg^*_t(x_i)}{D^*_t(S^*_t)}\right].
\end{aligned}$$

Let $\mathcal{A}$ the event $|D^*_t(S^*) - \frac{2mp}{1+p}t| < (C_N+1)t^{1/2}(\ln t)^2$ then

$$\begin{aligned}
&\mathbf{E}\left[\frac{deg^*_t(x_i)}{D^*_t(S^*_t)}\right] \\
&\quad = \mathbf{E}\left[\frac{deg^*_t(x_i)}{D^*_t(S^*_t)} \,\middle|\, \mathcal{A}\right]\mathbf{Pr}(\mathcal{A}) \\
&\qquad + \mathbf{E}\left[\frac{deg^*_t(x_i)}{D^*_t(S^*_t)} \,\middle|\, \quad \mathcal{A}\right]\mathbf{Pr}(\quad \mathcal{A}) \\
&\quad = \mathbf{E}\left[deg^*_t(x_i) \mid \mathcal{A}\right]\left(\frac{1+p}{2mpt} + \tilde{O}\left(t^{-3/2}\right)\right)\mathbf{Pr}(\mathcal{A}) \\
&\qquad + O\left(\mathbf{Pr}(\neg\mathcal{A})\right) \\
&\quad = \mathbf{E}\left[deg^*_t(x_i)\right]\left(\frac{1+p}{2mpt}\right) + \tilde{O}\left(t^{-1/2}\right) + O\left(\mathbf{Pr}(\neg\mathcal{A})\right) \\
&\quad = \mathbf{E}\left[deg^*_t(x_i)\right]\left(\frac{1+p}{2mpt}\right) + \tilde{O}\left(t^{-1/2}\right)
\end{aligned}$$

where we used the fact $\deg^*_t(x_i) \leq D^*_t(S^*_t) \leq 2mt$, and Lemma 3.5.

Therefore

$$\mathbf{E}\left[\deg^*_{t+1}(x_i)\right] = \mathbf{E}\left[\deg^*_t(x_i)\right]\left(1 + \frac{1}{t}\right) + \tilde{O}\left(t^{-1/2}\right),$$

and by induction

$$\mathbf{E}\left[\deg^*_t(x_i)\right] = \mathbf{E}\left[\deg^*_N(x_i)\right]t/N + \tilde{O}\left(t^{1/2}\right)$$

Now, if $t < N$ we have

$$\begin{aligned}\mathbf{E}\left[\deg^*_{t+1}(x_i)|G^*_t\right] &= \deg^*_t(x_i) + m\frac{\deg^*_t(x_i)}{2mt} \\ &= \deg^*_t(x_i)\left(1+\frac{1}{2t}\right).\end{aligned}$$

And therefore

$$\begin{aligned}\mathbf{E}\left[\deg^*_N(x_i)\right] &= \mathbf{E}\left[\deg^*_i x_i\right]\prod_{1\le j<i}\left(1+\frac{1}{2j}\right) \\ &= m\prod_{1\le j<i}\left(1+\frac{1}{2j}\right)\end{aligned}$$

□

Lemma 3.3. *Suppose $m \ge 4$. Let*

$$\epsilon_t = \mathbf{Pr}\left[\exists\tau \ge t:\ m^*_\tau - M^*_\tau \le m\right].$$

Then $\epsilon_t \to 0$ as $t \to \infty$.

Proof From Lemma 3.6,

$$\mathbf{Pr}\left[m^*_\tau < (2pm\tau)^{q/2+p/4}\right] = O\left(\tau^{-\frac{2+3p}{4}(m-1)}\right),$$

So for some constant $A > 0$ we have

$$\text{(3.6)}\quad \mathbf{Pr}\left[\exists\tau \ge t:\ m^*_\tau < (2pm\tau)^{q/2+p/4}\right] \le A\sum_{\tau\ge t}\tau^{-\frac{2+3p}{4}(m-1)} = O(t^{-\frac{2+3p}{4}(m-1)}).$$

Also, from Lemma 3.7,

$$\mathbf{Pr}\left[M^*_\tau \ge \tau^{q/2}(\ln\tau)^3\right] \le \exp\left(m - \frac{(\ln\tau)^2}{6}\right),$$

therefore

$$\begin{aligned}&\mathbf{Pr}\left[\exists\tau \ge t:\ M^*_\tau \ge \tau^{q/2}\ln(t)^3\right] \\ \text{(3.7)}\qquad &\le \sum_{\tau\ge t}\exp\left(m - \tfrac{(\ln\tau)^2}{6}\right) \\ &= O(e^{-(\ln t)^2/12}).\end{aligned}$$

The result follows from (3.6) and (3.7). □

Lemma 3.4. *Suppose $t \ge N$. Then*

$$\frac{2mp}{1+p}t \le \mathbf{E}\left[D^*_t(S^*)\right] \le \frac{2mp}{1+p}t + C_N t^{\frac{q}{2}}$$

Proof Let $z_t = \mathbf{E}\left[D^*_t(S^*)\right]$, then $z_N = 2Nm$,

$$z_{t+1} = z_t + mp + qm\frac{z_t}{2mt} = mp + z_t\left(1+\frac{q}{2t}\right).$$

The result follows by induction. □

Lemma 3.5. *If $t \ge N$ then*

$$\mathbf{Pr}\left[\left|D^*_t(S^*) - \tfrac{2mp}{1+p}t\right| \ge (C_N+1)t^{1/2}\ln t\right] \le 2e^{-p(\ln t)^2/m}.$$

Proof Enumerate the edges $e_1, e_2, \ldots, e_{mt}$ in the order they appear. For $i > Nm$ let Y_i be the $0,1$ random variable taking value 1 if and only if e_i is incident to S^*. Then

$$D^*_t(S^*) = 2Nm + \sum_{i=mN+1}^{mt} Y_i$$

and

$$\mathbf{Pr}\left[Y_i = 0 \mid D^*_t(S^*)\right] = q\left(1 - \frac{D^*_t(S^*)}{2m\lfloor i/m\rfloor}\right).$$

We apply Azuma's inequality to show the concentration of $D^*_t(S^*)$. Given i we define for $\tau = \lfloor i/m\rfloor + 1, \ldots, t$.

$$\begin{aligned}\Delta_\tau(i) = \Big|\ &\mathbf{E}\left[D^*_\tau(S^*)|Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}, Y_i = 0\right] \\ &- \mathbf{E}\left[D^*_\tau(S^*)|Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}, Y_i = 1\right]\Big|,\end{aligned}$$

Notice that

$$\Delta_{\tau+1}(i) = \Delta_\tau(i) + q\frac{\Delta_\tau(i)}{2m\lfloor t/m\rfloor},$$

and $\Delta_{\lfloor i/m\rfloor+1}(i) = 1$. Thus,

$$\Delta_\tau(i) \le \left(\frac{m\tau}{i}\right)^{q/2}.$$

Therefore,

$$\begin{aligned}\sum_{i=Nm+1}^{mt}\Delta_t(i)^2 &\le \sum_{i=Nm+1}^{mt}\left(\frac{mt}{i}\right)^q \\ &\le (mt)^q\int_{mN}^{mt} x^{-q}dx \le mt/p,\end{aligned}$$

and

$$\mathbf{Pr}\left[\left|D^*_t(S^*) - \mathbf{E}\left[D^*_t(S^*)\right]\right| \ge t^{1/2}\ln t\right] \le 2e^{\frac{-p(\ln t)^2}{m}}.$$

The result follows after using Lemma 3.4. □

Lemma 3.6. *If $i \leq N$ and $\epsilon > 0$ then*

$$\mathbf{Pr}\left[\deg_t^*(x_i) < (2pmt)^{1-\epsilon}\right] = O\left(t^{-\epsilon(m-1)}\right)$$

Proof We couple our graph process with an urn process: We start the process at time $t = N$ with $r = \deg_N^*(x_i)$ red balls and $b = 2Nm - r$ blue balls. Each time we add an edge to the graph that is incident to S^* we add a ball to the urn. If the edge is incident to x_i, the ball is red otherwise is blue. Then R_t the number of red balls in the urn by time t is equal to $\deg_t^*(x_i)$, while the total number of balls in the urns is $D_t^*(S^*)$.

Note that preferential attachment is equivalent to choosing an edge e at random and then choosing a random end point from e, therefore this urn process follows a Polya urn process: In time t given that we add a ball, the probability of adding a red ball is R_t/T_t, where T_t is the total number of balls in the urns. We think in our urn process isolated from the graph process and call "a step" of the process when a ball is added. We use $s = 1, 2, \ldots, D_t^*(S^*) - 2Nm$ to index the steps of the urn process.

Now, for any $0 \leq k \leq s$

$$\begin{aligned}
&\mathbf{Pr}\left[R_s = r + k\right] \\
&= \binom{s}{k} \frac{r \cdots (r+k-1) b(b+1) \cdots (b+s-k-1)}{(r+b) \cdots (r+b+s-1)} \\
&= \frac{(r+b-1)!}{(s+r)(r-1)!(b-1)!} \prod_{i=1}^{r-1} \frac{k+i}{s+i} \\
&\quad \cdot \prod_{i=1}^{r+k} \left(1 - \frac{b-1}{b+s-k+i-1}\right) \\
&\leq \frac{(r+b-1)!}{(s+r)(r-1)!(b-1)!} \left(\frac{k+r-1}{s+r-1}\right)^{r-1} \\
&\quad \left(1 - \frac{b-1}{b+s+r-1}\right)^{r+k}
\end{aligned}$$

And therefore if $\epsilon > 0$

$$\begin{aligned}
&\mathbf{Pr}\left[R_s \leq s^{1-\epsilon}\right] \\
&\leq \frac{(r+b-1)!}{(s+r)(r-1)!(b-1)!} \sum_{k=0}^{s^{1-\epsilon}-r} \left(\frac{k+r-1}{s+r-1}\right)^{r-1} \\
&\leq \frac{(r+b-1)!}{(r-1)!(b-1)!} \int_0^{s^{-\epsilon}} x^{r-1} dx \\
&\leq \frac{2^{r+b}}{r-1} s^{-\epsilon(r-1)}
\end{aligned}$$

Recalling that $r \geq m$ and $r + b = 2Nm$ and $\deg_t^*(x_i) = R_{D_t^*(S^*)-2Nm}$ we get, using Lemma 3.5,

$$\begin{aligned}
&\mathbf{Pr}\left[\deg_t^*(x_i) \leq (2pmt)^{1-\epsilon}\right] \\
&\leq \mathbf{Pr}\left[\deg_t^*(x_i) \leq t^{1-\epsilon} | D_t^*(S^*) - 2Nm \geq 2pmt\right] \\
&\qquad + \mathbf{Pr}\left[D_t^*(S^*) - 2Nm < 2pmt\right] \\
&\leq \mathbf{Pr}\left[R_s \leq s^{1-\epsilon} | s \geq 2pmt\right] + e^{-p(\ln t)^2/m} \\
&\leq 2^{mN}(2pmt)^{-\epsilon(m-1)} + e^{-p(\ln t)^2/m} \\
&= O\left(t^{-\epsilon(m-1)}\right).
\end{aligned}$$

□

Lemma 3.7. *Let $s > N$ and let $t \geq s$.*

$$\mathbf{Pr}\left[\deg_t^*(x_s) \geq (t/s)^{q/2}(\ln t)^3\right] \leq \exp\left(m - \frac{(\ln t)^2}{6}\right)$$

Proof Fix $s > N$ and let $X_\tau = \deg_\tau^*(s)$ for $\tau = s, s+1, \ldots, t$.

Then conditional on $X_\tau = x$, we have

$$X_{\tau+1} = X_\tau + B\left(m, \frac{qx}{2m\tau}\right) \tag{3.8}$$

and so

$$\begin{aligned}
\mathbf{E}\left[e^{\lambda X_{\tau+1}} \mid X_\tau = x\right] &= e^{\lambda x}\left(1 - \frac{qx}{2m\tau} + \frac{qx}{2m\tau}e^\lambda\right)^m \\
&\leq e^{\lambda x} \exp\left(\frac{qx}{2\tau}(e^\lambda - 1)\right) \\
&= \exp\left(\lambda x\left(1 + q\frac{(1+\lambda)}{2\tau}\right)\right),
\end{aligned}$$

for any $\lambda \leq 1$.

Thus

$$\mathbf{E}\left[e^{\lambda X_{\tau+1}}\right] \leq \mathbf{E}\left[\exp\left(X_\tau \lambda\left(1 + \frac{q(1+\lambda)}{2\tau}\right)\right)\right].$$

If we put $\lambda_{\tau-1} = \lambda_\tau\left(1 + \frac{q(1+\lambda_\tau)}{2\tau}\right)$ and take $\lambda_t = \lambda$ small enough such that

$$\lambda_\tau \leq \Lambda = \min\left\{1, \frac{1}{\ln(t/s)}\right\} \text{ for } \tau = s, \ldots, t, \tag{3.9}$$

we have

$$\mathbf{E}(e^{\lambda X_t}) \leq e^{m\lambda_s}.$$

and we can write

$$\lambda_{\tau-1} \leq \lambda_\tau\left(1 + \frac{(1+\Lambda)q}{2\tau}\right),$$

then

$$\begin{aligned}
\lambda_s &\leq \lambda \prod_{\tau=s}^{t}\left(1 + \frac{(1+\Lambda)q}{2\tau}\right) \\
&\leq 2\lambda(t/s)^{(1+\Lambda)q/2} \\
&\leq 6\lambda(t/s)^{q/2}
\end{aligned}$$

and therefore we can take $\lambda = \frac{\Lambda}{6}(s/t)^{q/2}$ and get (3.9).

Putting $u = (t/s)^{q/2}(\ln t)^3$ we get

$$\begin{aligned}\mathbf{Pr}(X_t \geq u) &\leq e^{m\lambda_s - \lambda u} \\ &\leq \exp\left(\Lambda m - \frac{\Lambda(\ln t)^3}{6}\right) \\ &\leq \exp\left(m - \frac{(\ln t)^2}{6}\right)\end{aligned}$$

□

References

[1] A. Barabasi and R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509-512.

[2] J. Cho and S. Roy, Impact of search engines on page popularity.

[3] C. Cooper and A. M. Frieze, A General Model of Undirected Web Graphs, Random Structures and Algorithms 22 (2003) 311-335

[4] E. Drinea, A.M. Frieze and M. Mitzenmacher, Balls and Bins Models with Feedback, Proceedings of SODA 2002, 308-315.

[5] A. Flaxman, A.M. Frieze and T.I. Fenner, High degree vertices and eigenvalues in the preferential attachment graph, Proceedings of RANDOM 2003.

[6] M. Hindman, K. Tsioutsiouliklis and J. A Johnson, Googlearchy: How a Few Heavily-Linked Sites Dominate Politics on the Web, Annual Meeting of the Midwest Political Science Association, 2003.

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

[8] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal. Stochastic models for the web-graph. Proc. 41st Annual Symp on Foundations of Computer Science, 2000.

[9] N.L. Johnson and S. Kotz, Urn models and their application : an approach to modern discrete probability theory, Wiley, New York, 1977.