The influence of search engines on preferential attachment

Alan Frieze[*] Juan Vera[†] Soumen Chakrabarti[†]

Abstract

There is much current interest in the evolution of social networks, in particular, the Web graph, through time. "Preferential attachment" and the "copying model" are well-known models which explain the observed power-law degree distribution of the Web graph reasonably well. However, existing evolution models do not include the significant influence of search engines on how Web page authors find existing pages and create links to them. Recent applied work has raised the concern that highly popular search engines limit the attention of authors to a small set of "celebrity" URLs, for any query. Page authors 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. Barabási and Albert [2] were among the first to propose a generative model of the Web, called *preferential attachment*, which leads to a power-law distribution $\mathbf{Pr}(\text{degree} = k) \propto k^{-3}$, i.e., a power of 3. Introducing random links some fraction of the time allowed Pennock et al [14] to bring the power closer to empirically observed data (2.1 for in-degree and 2.38 for out-degree).

Preferential attachment explains a power-law degree distribution but not the presence of a large number of bipartite cliques in the Web graph. Kleinberg *et al.* [11] 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.* [12] analyzed the copying process to show that it, too, leads to a power law degree distribution with a power of approximately 2, which is closer to empirical observation. The copying model naturally explains the presence of bipartite cliques.

Both the preferential attachment and the copying model assume organic evolution of the Web graph, without any powerful central entity influencing a large number of Web page authors with

[*]Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, USA. Research supported in part by NSF grant CCR-0200945.

[†]Indian Institute of Technology, Bombay, India. Research done while visiting the Department of Computer Science, Carnegie Mellon University

regard to how they link to existing pages. This is precisely the role that search engines like Google or Yahoo! fulfill. Web page authors learn about a topic by launching queries to a search engine. The search engine responds with a limited number (10–20) of hits per page, and users rarely foray beyond the first 1–2 pages of hits [4].

With Web search proliferating to over five billion searches per month[1] as of February 2006, many Web page authors presumably create links to pages that they found by using a search engine. In other words, with some probability every link in the Web graph owes its existence to a search engine, and therefore, the evolution of the Web graph has been influenced permanently and pervasively by the existence of search engines.

In their early days, search engines merely observed and exploited the Web graph (specifically, links to a page) for ranking. Now, they are unquestionably influencing the evolution of Web graph as well. Most search engines today pay attention to in-degree and Pagerank [3] while ranking results. This can potentially set up a "virtuous circle of limelight": a search engine ranks a page highly, authors find the page more often, some of them link to it, raising its in-degree and Pagerank, which leads to a further improvement or entrenchment of its rank.

The virtuous circle can be brutal to new pages and sites: Cho and Roy [4] 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. In a theoretical setting, Drinea *et al.* [6] analyze a balls-and-bins process with a related feedback mechanism, and show that positive feedback leads to a rapid landslide victory for the winning bin. Pandey *et al.* [13] confirms that introducing some randomness in the ranking function creates a better exploration-exploitation trade-off, avoiding the worst effects of the virtuous circle.

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 the power-law form derived in earlier work.

For simplicity, like Barabási *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 (multi)-graphs $G_t, t = 1, 2, \ldots$. 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:

[1] `http://www.nielsen-netratings.com/pr/PR_033006_UK.pdf`

$\deg_t(x)$ 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 $S_t = V_t$.)

z_t is the smallest degree of vertices in S_t

Z_t is the largest degree of vertices in $V_t \setminus S_t$

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

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

We use process $\mathcal{P}$ to generate the graph sequence $G_t = (V_t, E_t)$, for $t = 1, 2, \ldots, n$:

Formal Definition of Process $\mathcal{P}$

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:

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

$\diamond\diamond$ 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.

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. Also, as p increases, the power (negative slope) increases as well: at $p = 0$ it is 2.8, at $p = 0.3$ it is 3.96, and at $p = 0.6$ it is 5.9.

Furthermore, as Figure 2(a) 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 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 2(b) 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.

1.2 Our results

We will prove the following, where all asymptotic notation is with respect to n, the number of steps for which the process $\mathcal{P}$ is run (which is the same as the number of nodes).

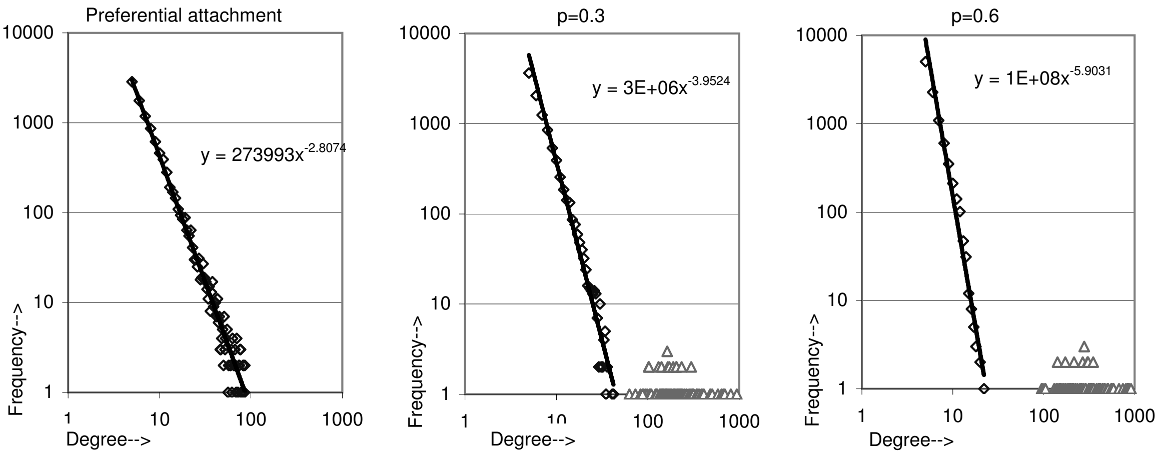


Figure 1: The presence of a search engine in our model makes the power in the degree power law more negative, and, with increasing p, separates out the celebrities from the non-celebrities ($N = 100$, $|V| = 10000$, and $m = 5$).

Theorem 1. *Let* $m \geq \max\{15, 2/q\}$ *and* $0 < p < 1$.

(a) *Let* $S_n = \{s_1, \ldots, s_N\}$ *in decreasing order of degree. Then* $\mathbf{E}[\deg_n(s_i)] \sim \alpha_i n$ *for every* $i \leq N$, *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* $k \geq m$,

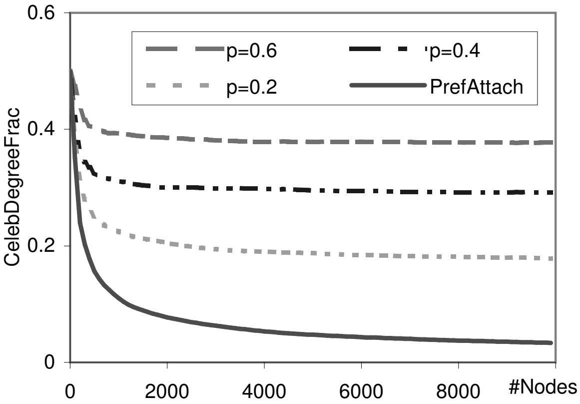
$$\begin{aligned} \overline{d}_k(n) &= \frac{2n}{2+mq} \prod_{i=m+1}^{k} \frac{i-1}{i+2/q} + \tilde{O}(n^{q/2}) \\ &= \frac{A_1 n}{k^{1+2/q}} + \tilde{O}(n/k^{2+2/q} + n^{q/2}). \end{aligned}$$

The theorem and its proof confirms all the features we see in the simulations: celebrities capture a large fraction of links, the celebrity list gets fixed quickly, and non-celebrities follow a power-law degree distribution with a power steeper than in preferential attachment.

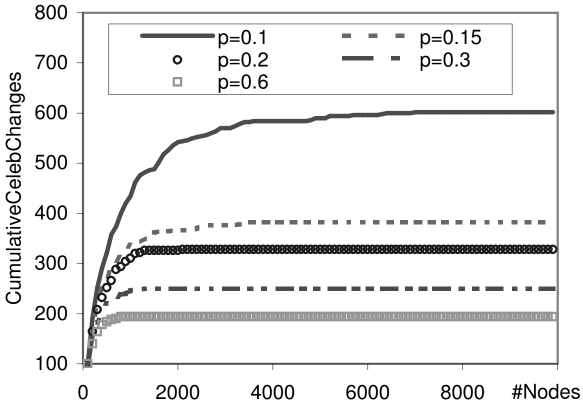
The proof depends on showing that after time t_0, the first time there is a considerable degree gap between the celebrity list and the non-celebrities the probability of having a non-celebrity move to the celebrity list is very small. So in practice the celebrity list becomes fixed. Once the celebrity list is fixed, our process $\mathcal{P}$ looks very similar to an analogous process $\mathcal{P}^*$ where in each step S_t is replaced by $S_t^* = S^* = \{x_1^*, \ldots, x_N^*\}$ in decreasing order of degree (if $t < N$ take $S_t^* = V_t$). In other words, in process $\mathcal{P}^*$ we take the N *oldest* vertices as S_t^*, instead of the N largest-degree vertices.

Let G_t be the sequence of graphs produced by process $\mathcal{P}$ and G_t^* be the sequence of graphs produced by process $\mathcal{P}^*$. In Section 2 we construct a coupling for $t = 1, 2, \ldots$ between the sequence of graphs G_t and the sequence of graphs G_t^*.

In Section 3 we compute $\mathbf{E}[\deg_n(s)]$ for $s \in S_n$ and $\mathbf{E}[d_k(n)]$, conditioning on t_0 and G_{t_0}. In Section 4 we get bounds for the probability of having a small gap between celebrities and non-celebrities. Finally, in Section 5 we give the proof of Theorem 1.



(a) $N = 10$, $n = 10000$, and $m = 5$



(b) $N = 100$, $n = 10000$, and $m = 5$

Figure 2: (a) 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. (b) 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.

2 Coupling G_t and G_t^*

Let z_t be the degree of the lowest degree vertex in S_t and Z_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 z_t and Z_t and then from this time on S_t remains fixed. In this sense the graph G_t is very similar to the graph $G_t^* = (V_t^*, E_t^*)$, constructed by process $\mathcal{P}^*$, where the top N is fixed from the beginning (the top is fixed by age not by degree). We define z_t^* and Z_t^* in G_t^* in an analogous way to z_t and Z_t.

Lemma 1. *We can couple $\mathcal{P}$ and $\mathcal{P}^*$ in such a way that for all $t > 0$,*

$$D_t^*(S_t) \leq D(S_t), \quad z_t^* \leq z_t \quad \text{and} \quad Z_t^* \geq Z_t.$$

Proof To couple processes $\mathcal{P}$ and $\mathcal{P}^*$, we first define $\mathcal{P}'$, a small modification of process $\mathcal{P}$. If after time step t vertex $v \in V_t' \setminus S_t'$ has degree larger than the degree of u (the minimum-degree vertex in S_t'), then, instead of moving v into S_t' and u out of S_t', we change the endpoint of some of the last edges inserted incident to v to make them incident to u (leaving the other end fixed), in order to swap the degree of v and u. In this way the degree sequence in maintained, and S_t' remains fixed as the N oldest vertices (i.e. $S_t' = S_t^*$). The graph generated by $\mathcal{P}'$ has a different edge structure to G_t, but it has the same degree sequence, thus

$$D_t'(S_t') = D(S_t), \quad z_t' = z_t \quad \text{and} \quad Z_t' = Z_t. \tag{1}$$

Now let $\mathcal{P}''$ be the process where after each step we proceed almost the same way as in $\mathcal{P}'$, except that in $\mathcal{P}'$ if some k endpoints are changed (from a vertex outside S_t' to a vertex in S_t'), in $\mathcal{P}''$ we don't make these changes, but instead we move the endpoints of k random edges chosen uniformly from the last inserted edges incident to $V_t \setminus S_t''$, to k random positions chosen by preferential attachment in $S_t'' = S_t' = S_t^*$.

We can think of every edge inserted as two directed edges, and then choosing by preferential attachment is equivalent to choosing a random edge uniformly and then choosing its destination vertex. This permits us to couple process $\mathcal{P}'$ and $\mathcal{P}''$ in such way that

$$D''_t(S''_t) = D'(S'_t), \quad z''_t \leq z'_t \quad \text{and} \quad Z''_t = Z'_t. \tag{2}$$

Notice that process $\mathcal{P}''$ looks like process $\mathcal{P}^*$ except that the probability of applying step $\diamond^*$ is greater than p (and depends on G_t). So we can couple $\mathcal{P}^*$ and $\mathcal{P}''$ in such way that

$$D^*_t(S_t) \leq D''(S''_t), \quad z^*_t \leq z''_t \quad \text{and} \quad Z^*_t \geq Z''_t. \tag{3}$$

Putting equations (1), (2) and (3) together we get that we can couple $\mathcal{P}$ and $\mathcal{P}^*$ such that

$$D^*_t(S_t) \leq D(S_t), \quad z^*_t \leq z_t \quad \text{and} \quad Z^*_t \geq Z_t. \tag{4}$$

□

3 Analysis of the degree distribution of G^*_t

Given a time $t > 0$, let $\mathcal{G}_t$ be the set of graphs with vertex set $\{x_1, \ldots, x_t\}$ and mt edges. Notice that $G_t, G^*_t \in \mathcal{G}_t$. In this section we analyze the behavior of G^*_t when process $\mathcal{P}^*$ is initialized at time t with some graph $G \in \mathcal{G}_t$, i.e., when conditioning on $G^*_t = G$. In Lemma 2 we prove that $\overline{d^*_k}(n)$ follows a power law, while in Lemma 3 we prove that the expected $\deg^*_n(x^*_i)$ grows linearly with time for nodes x^*_i with $i \leq N$.

Lemma 2. *Fix a time $t_0 \geq N$ and let $G \in \mathcal{G}_{t_0}$. Then, at a later time $n \geq t_0$, for all $k \geq m$*

$$\begin{aligned}
\mathbf{E}\left[d^*_k(n) \mid G^*_{t_0} = G\right] &= \frac{2n}{2+mq} \prod_{i=m+1}^{k} \frac{i-1}{i+2/q} + \tilde{O}(t_0 + n^{q/2}) \\
&= \frac{A_1 n}{k^{1+2/q}} + \tilde{O}(t_0 + n^{q/2} + n/k^{2+2/q}).
\end{aligned}$$

Proof Our approach to proving a power law is to find a recurrence for $\mathbf{E}\left[d^*_k(n) \mid G^*_{t_0} = G\right]$.

Notice that $\mathbf{E}\left[d^*_{m-1}(\tau) \mid G^*_{t_0} = G\right] = 0$ for all $\tau > 0$. Then for $\tau \geq t_0, k \geq m$,

$$\begin{aligned}
\mathbf{E}\left[d^*_k(\tau+1) \mid G^*_\tau, G^*_{t_0}\right] &= d^*_k(\tau) + qm\left(\frac{(k-1)d^*_{k-1}(\tau)}{2m\tau} - \frac{kd^*_k(\tau)}{2m\tau}\right) + 1_{k=m} + O(Z^*_\tau \tau^{-1}) \qquad (5) \\
&= d^*_k(\tau) + q\frac{(k-1)d^*_{k-1}(\tau) - kd^*_k(\tau)}{2\tau} + 1_{k=m} + O(Z^*_\tau \tau^{-1}).
\end{aligned}$$

Explanation of (5)**:** qm is the expected number of edges involving non-celebrities. The expression following qm is the probability that an additional edge convertes a vertex of degree $k-1$ to one of degree k less the probability that it converts a vertex of degree k into one of degree $k+1$. The $O(Z^*_\tau \tau^{-1})$ term accounts for the addition of parallel edges.

Taking expectations w.r.t. G^*_τ, we get

$$\begin{aligned}
&\mathbf{E}\left[d^*_k(\tau+1) \mid G^*_{t_0} = G\right] = \mathbf{E}\left[d^*_k(\tau) \mid G^*_{t_0} = G\right] \\
&+ q\frac{(k-1)\mathbf{E}\left[d^*_{k-1}(\tau) \mid G^*_{t_0} = G\right] - k\mathbf{E}\left[d^*_k(\tau) \mid G^*_{t_0} = G\right]}{2\tau} + 1_{k=m} + O(\mathbf{E}\left[Z^*_\tau \mid G^*_{t_0} = G\right]\tau^{-1}). \qquad (6)
\end{aligned}$$

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

$$f_k = 1_{k=m} + q\frac{(k-1)f_{k-1}-kf_k}{2} \tag{7}$$

yielding

$$f_m = \frac{2}{2+mq}$$

and

$$\begin{aligned} f_k &= \frac{2}{2+mq}\prod_{i=m+1}^{k}\frac{i-1}{i+2/q} \\ &= \frac{2}{2+mq}\exp\left(\sum_{i=m+1}^{k}\ln\left(\frac{i-1}{i+2/q}\right)\right) \\ &= \frac{2e^{-g(m,k)}}{2+mq}\left(\frac{m+1+2/q}{k+2/q}\right)^{1+2/q} \\ &= \frac{2e^{-g(m)}(m+1+2/q)^{1+2/q}}{2+mq}\frac{e^{O(1/k)}}{k^{1+2/q}} \end{aligned}$$

where

$$g(m,k)=\left(\sum_{i=m+1}^{k}\frac{1+2/q}{i+2/q}-\int_{x=m+1}^{k}\frac{1+2/q}{x+2/q}dx\right)+\sum_{i=m+1}^{k}\sum_{l=2}^{\infty}\frac{1}{l}\left(\frac{1+2/q}{i+2/q}\right)^{l}=g_m+O(k^{-1})$$

where $g_m=\lim_{k\to\infty}g(m,k)$.

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

$$|\mathbf{E}\left[d_k^*(\tau)\mid G_{t_0}^*=G\right]-f_k\tau|\leq M(t_0+\tau^{q/2}(\ln\tau)^3) \tag{8}$$

for all $\tau>0$.

Let

$$\Theta_k(\tau)=\mathbf{E}\left[d_k^*(\tau)\mid G_{t_0}^*=G\right]-f_k\tau.$$

Lemma 8 (proved later) implies that $\mathbf{E}\left[Z_\tau^*\mid G_{t_0}^*=G\right]\leq O\left(\tau^{q/2}(\ln\tau)^3\right)$. So after taking expectations over G_τ^* in (5) and substituting $\mathbf{E}\left[d_k^*(\tau)\mid G_{t_0}^*=G\right]=\Theta_k(\tau)+f_k\tau$ we see that for $k\geq m$ and $\tau\geq t_0$,

$$\begin{aligned} &f_k(\tau+1)+\Theta_k(\tau+1)= \\ &\qquad f_k\tau+\Theta_k(\tau)+q\frac{(k-1)(\Theta_{k-1}(\tau)+f_{k-1}\tau)-k(\Theta_k(\tau)+f_k\tau)}{2\tau}+1_{k=m}+O(\tau^{q/2-1}(\ln\tau)^3). \end{aligned}$$

Using (7) to eliminate the f_k's we obtain

$$\Theta_k(\tau+1)=\left(1-\frac{qk}{2\tau}\right)\Theta_k(\tau)+\frac{q(k-1)}{2\tau}\Theta_{k-1}(\tau)+O(\tau^{q/2-1}(\ln\tau)^3). \tag{9}$$

Let L denote the hidden constant in $O(\tau^{q/2-1}(\ln\tau)^3)$ of (9). Our inductive hypothesis $\mathcal{H}_\tau$ is that $|\Theta_k(\tau)| \le M(t_0+\tau^{q/2}(\ln\tau)^3)$ for every $k\ge m$. It is trivially true for $\tau\le t_0$. So assume that $\tau\ge t_0$. Then, from (9),

$$\begin{aligned}|\Theta_k(\tau+1)| &\le M(t_0+\tau^{q/2}(\ln\tau)^3)+L\tau^{q/2-1}(\ln\tau)^3\\ &\le M(t_0+(\tau+1)^{q/2}(\ln\tau)^3)\end{aligned}$$

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

Lemma 3. *Fix a time $t_0\ge N$ and $G\in\mathcal{G}_{t_0}$. Then, at a later time $n\ge t_0$, for all $i\le N$,*

$$\mathbf{E}\left[\deg_n^*(x_i^*)\mid G_{t_0}^*=G\right]=n\frac{\deg_G(x_i^*)}{t_0}+\tilde{O}\left((n/t_0)^{5/6}\right)+O\left(t_0^{3/2}\right).$$

Proof Notice that from Lemma 5 (below), if $\tau>t_0$,

$$\begin{aligned}\left|\mathbf{E}\left[D_\tau^*(S_\tau^*)|G_{t_0}^*=G\right]-\frac{2mp}{1+p}\tau\right|>2e^{q/2}\tau^{5/6} &\Rightarrow \left|D(S^*(G))-\frac{2mp}{1+p}t_0\right|>2\tau^{(5-3q)/6}t_0^{q/2}\\ &\Rightarrow 2mt_0>2\tau^{(5-3q)/6}t_0^{q/2}\\ &\Rightarrow m^3t_0^{3/2}>\tau,\end{aligned}\tag{10}$$

where we used $D(S_G^*)\le 2mt_0$.

Now, let $\mathcal{A}_\tau^*$ be the event

$$\left|D_\tau^*(S_\tau^*)-\frac{2mp}{1+p}\tau\right|<4e\tau^{5/6}.$$

Then

$$\begin{aligned}\mathbf{Pr}\left[\neg\mathcal{A}_\tau^*|G_{t_0}^*=G\right] &\le \mathbf{Pr}\left[\left|D_\tau^*(S_\tau^*)-\mathbf{E}\left[D_\tau^*(S_\tau^*)|G_{t_0}^*=G\right]\right|\ge 2e\tau^{5/6}\,\middle|\,G_{t_0}^*=G\right]\\ &\quad+1_{\left|\mathbf{E}\left[D_\tau^*(S_\tau^*)|G_{t_0}^*=G\right]-\frac{2mp}{1+p}\tau\right|>2e\tau^{5/6}}\\ &\le 2e^{-p(\ln\tau)^2/8m^3}+1_{m^3t_0^{3/2}>\tau}\end{aligned}\tag{11}$$

where the last line follows from Lemma 6 (below) and (10).

If $\tau\ge N$, then

$$\mathbf{E}\left[\deg_{\tau+1}^*(x_i^*)|G_\tau^*,G_{t_0}^*=G\right]=\deg_\tau^*(x_i^*)+mq\frac{\deg_\tau^*(x_i^*)}{2m\tau}+mp\frac{\deg_\tau^*(x_i^*)}{D_\tau^*(S_\tau^*)}.$$

Taking expectations w.r.t. G_τ^*, in the conditional space $G_{t_0}^*=G$, we get for every $\tau\ge t_0$

$$\mathbf{E}\left[\deg_{\tau+1}^*(x_i^*)\mid G_{t_0}^*=G\right]=\mathbf{E}\left[\deg_\tau^*(x_i^*)\mid G_{t_0}^*=G\right]\left(1+\frac{q}{2\tau}\right)+mp\mathbf{E}\left[\frac{\deg_\tau^*(x_i^*)}{D_\tau^*(S_\tau^*)}\,\middle|\,G_{t_0}^*=G\right].$$

But,

$$
\begin{aligned}
\mathbf{E}\left[\frac{\deg^*_\tau(x^*_i)}{D^*_\tau(S^*_\tau)} \,\middle|\, G^*_{t_0}=G\right] &= \mathbf{E}\left[\frac{\deg^*_\tau(x^*_i)}{D^*_\tau(S^*_\tau)} \,\middle|\, \mathcal{A}^*_\tau, G^*_{t_0}=G\right]\mathbf{Pr}(\mathcal{A}^*_\tau \mid G^*_{t_0}=G) \\
&\quad + \mathbf{E}\left[\frac{\deg^*_\tau(x^*_i)}{D^*_\tau(S^*_\tau)} \,\middle|\, \neg\mathcal{A}^*_\tau, G^*_{t_0}=G\right]\mathbf{Pr}(\neg\mathcal{A}^*_\tau \mid G^*_{t_0}=G) \\
&= \mathbf{E}\left[\deg^*_\tau(x^*_i) \mid \mathcal{A}^*_\tau, G^*_{t_0}=G\right]\left(\frac{1+p}{2mp\tau}+\tilde{O}\left(\tau^{-7/6}\right)\right)\mathbf{Pr}(\mathcal{A}^*_\tau \mid G^*_{t_0}=G) \\
&\quad + O\left(\mathbf{Pr}(\neg\mathcal{A}^*_\tau \mid G^*_{t_0}=G)\right) \\
&= \mathbf{E}\left[\deg^*_\tau(x^*_i) \mid G^*_{t_0}=G\right]\left(\frac{1+p}{2mp\tau}\right)+\tilde{O}\left(\tau^{-1/6}\right)+O\left(\mathbf{Pr}(\neg\mathcal{A}^*_\tau \mid G^*_{t_0}=G)\right) \\
&= \mathbf{E}\left[\deg^*_\tau(x^*_i) \mid G^*_{t_0}=G\right]\left(\frac{1+p}{2mp\tau}\right)+\tilde{O}\left(\tau^{-1/6}\right)+O\left(1_{m^3t_0^{3/2}>\tau}\right),
\end{aligned}
$$

where we used the fact $\deg^*_\tau(x^*_i) \le D^*_\tau(S^*_\tau) \le 2m\tau$ and (11). Therefore

$$
\mathbf{E}\left[\deg^*_{\tau+1}(x^*_i) \mid G^*_{t_0}=G\right] = \mathbf{E}\left[\deg^*_\tau(x^*_i) \mid G^*_{t_0}=G\right]\left(1+\frac{1}{\tau}\right)+\tilde{O}\left(\tau^{-1/6}\right)+O\left(1_{m^3t_0^{3/2}>\tau}\right),
$$

and by induction, for every $n > t_0$,

$$
\mathbf{E}\left[\deg^*_n(x^*_i) \mid G^*_{t_0}=G\right] = n\frac{\deg_G(x^*_i)}{t_0}+\tilde{O}\left((n/t_0)^{5/6}\right)+O\left(t_0^{3/2}\right)
$$

$\square$

Now we prove

Lemma 4. *There exists $D \ge 0$ such that the sequence*

$$
\frac{\deg_t(s_i)}{t}-\frac{D}{t^{1/6}}, \tag{12}
$$

$t \ge N$ is a sub-martingale.

Proof Proceeding as in Lemma 3, let $\mathcal{A}_\tau$ be the event

$$
\left|D_\tau(S_\tau)-\frac{2mp}{1+p}\tau\right| < 4e\tau^{5/6}.
$$

Then, if $\tau \ge m^3t_0^{3/2}$,

$$
\mathbf{E}\left[\deg_{\tau+1}(s_i) \mid G_\tau\right] \ge \deg_\tau(s_i)+mq\frac{\deg_\tau(s_i)}{2m\tau}+mp\frac{\deg_\tau(s_i)}{D_\tau(S_\tau)}.
$$

Taking expectations w.r.t. G_τ we obtain

$$
\mathbf{E}\left[\deg_{\tau+1}(s_i)\right] \ge \mathbf{E}\left[\deg_\tau(s_i)\right]\left(1+\frac{q}{2\tau}\right)+mp\mathbf{E}\left[\frac{\deg_\tau(s_i)}{D_\tau(S_\tau)}\right]. \tag{13}
$$

But,

$$\begin{aligned}
\mathbf{E}\left[\frac{\deg_\tau(s_i)}{D_\tau(S_\tau)}\right] &\geq \mathbf{E}\left[\frac{\deg_\tau(s_i)}{D_\tau(S_\tau)}\middle|\mathcal{A}_\tau\right]\mathbf{Pr}\left[\mathcal{A}_\tau\right]\\
&\geq \frac{1+p}{2mp\tau}\left(1-\frac{2e(1+p)}{mp\tau^{1/6}}\right)\mathbf{E}\left[\deg_\tau(s_i)\mid\mathcal{A}_\tau\right]\mathbf{Pr}(\mathcal{A}_\tau)\\
&\geq \frac{1+p}{2mp\tau}\left(1-\frac{2e(1+p)}{mp\tau^{1/6}}\right)\left(\mathbf{E}\left[\deg_\tau(s_i)\right]-2m\tau\mathbf{Pr}(\neg\mathcal{A}_\tau)\right)\\
&\geq \mathbf{E}\left[\deg_\tau(s_i)\right]\left(\frac{1+p}{2mp\tau}\right)-\frac{2e(1+p)^2}{mp^2\tau^{1/6}}-\frac{2}{p}e^{-\frac{p(\ln\tau)^2}{8m^3}},
\end{aligned}$$

where we used $\deg_\tau(s_i)\leq D_\tau(S_\tau)\leq 2m\tau$ and Lemma 6 together with Lemma 1.

Substituting into (13) we see that there is a constant $D'=D'(m,p)\geq 0$ such that for every $\tau\geq 1$,

$$\mathbf{E}\left[\frac{\deg_{\tau+1}(s_i)}{\tau+1}\right]\geq\mathbf{E}\left[\frac{\deg_\tau(s_i)}{\tau}\right]-\frac{D'}{\tau^{7/6}}.$$

(We may have to adjust the value of D' to account for small $\tau<m^3t_0^{3/2}$).

This implies that

$$\mathbf{E}\left[\frac{\deg_{\tau+1}(s_i)}{\tau+1}\right]-\frac{D'}{12(\tau+1)^{1/6}}\geq\mathbf{E}\left[\frac{\deg_\tau(s_i)}{\tau}\right]-\frac{D'}{12\tau^{1/6}}.$$

□

4 The celebrity list gets fixed

We first show that the total degree of celebrities, $D_t^*(S^*)$, is concentrated **whp**, and is, in expectation, a constant fraction of all edges ever added to the graph, as evident from simulation results shown in Figure 2. Then we show that z_t^* and Z_t^* are also concentrated **whp**. These results lead us to bounds on the probability of having a small gap between celebrities and non-celebrities.

Lemma 5. *Fix $t_0\geq N$ and $G\in\mathcal{G}_{t_0}$. Let $t\geq t_0$, then*

$$\left|\mathbf{E}\left[D_t^*(S^*)|G_{t_0}=G\right]-\frac{2mp}{1+p}t\right|\leq\left|D(S^*(G))-\frac{2mp}{1+p}t_0\right|\left(\frac{te}{t_0}\right)^{q/2}.$$

Proof Let $\tau\geq N$ then

$$\mathbf{E}\left[D_{\tau+1}^*(S^*)|D_\tau^*(S^*)\right]=D_\tau^*(S^*)+mp+qm\frac{D_\tau^*(S^*)}{2m\tau}=mp+D_\tau^*(S^*)\left(1+\frac{q}{2\tau}\right).$$

Thus,

$$\left|\mathbf{E}\left[D_{\tau+1}^*(S^*)|D_\tau^*(S^*)\right]-\frac{2mp}{1+p}(\tau+1)\right|=\left|D_\tau^*(S^*)-\frac{2mp}{1+p}\tau\right|\left(1+\frac{q}{2\tau}\right)$$

It follows that

$$\left|\mathbf{E}\left[D_t^*(S^*)|G_{t_0}=G\right]-\frac{2mp}{1+p}t\right|\leq\left|D(S^*(G))-\frac{2mp}{1+p}t_0\right|\exp\left\{\sum_{\tau=t_0}^{t-1}\frac{q}{2\tau}\right\}$$

and we observe that $\sum_{\tau=t_0}^{t-1}1/\tau\leq 1+\ln(t/t_0)$. □

Lemma 6. *Fix $t_0 \geq N$ and $G \in \mathcal{G}_{t_0}$. Let $\mathcal{B}$ denote the event $G^*_{t_0} = G$. For $t \geq t_0$ and $\lambda > 0$,*

$$\mathbf{Pr}\left[|D^*_t(S^*) - \mathbf{E}\left[D^*_t(S^*) \mid \mathcal{B}\right]| \geq \lambda t^{1/2} \ln t \mid \mathcal{B}\right] \leq 2e^{-\lambda^2 p (\ln t)^2 / 8m^3}.$$

Proof We condition on $\mathcal{B}$ and omit this explicit conditioning in our expressions. Enumerate the edges $e_1, e_2, \ldots, e_{mt}$ in the order they appear. For $i > t_0 m$ let Y_i be the $0,1$ random variable taking value 1 if and only if e_i is incident to S^*. Then

$$\begin{aligned} D^*_t(S^*) &= D(S^*) + \sum_{i=mt_0+1}^{mt} Y_i. \\ \mathbf{Pr}\left[Y_i = 0 \mid D^*_{\lfloor i/m \rfloor}(S^*)\right] &= q\left(1 - \frac{D^*_{\lfloor i/m \rfloor}(S^*)}{2m\lfloor i/m \rfloor}\right). \end{aligned}$$

We apply Azuma's inequality (see Alon and Spencer [1]) to show the concentration of $D^*_t(S^*)$. Given $t_0 m < i \leq tm$, fix $y_1, \ldots, y_{i-1} \in \{0,1\}$ and let

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

for $\tau = \lceil i/m \rceil, \ldots, t$.

Notice that

$$\Delta_{\tau+1}(i) = \Delta_\tau(i) + q\frac{m\Delta_\tau(i)}{2m\tau} \text{ and } \Delta_{\lceil i/m \rceil}(i) \leq m,$$

So

$$\begin{aligned} \Delta_\tau(i) &\leq m \prod_{j=\lceil i/m \rceil}^{\tau-1} \left(1 + \frac{q}{2j}\right) \\ &\leq 2m\left(\frac{m\tau}{i}\right)^{q/2}. \end{aligned}$$

Clearly $\Delta_\tau(I) \geq 0$ and therefore,

$$\sum_{i=t_0 m+1}^{mt} \Delta_t(i)^2 \leq 4m^2 \sum_{i=t_0 m+1}^{mt} \left(\frac{mt}{i}\right)^q \leq 4m^2 (mt)^q \int_{mt_0}^{mt} x^{-q} dx \leq 4m^3 t/p,$$

and the Lemma follows. □

Lemma 7. *If $i \leq N$ and $0 < A \ll t$ then*

$$\mathbf{Pr}\left[\deg^*_t(x^*_i) < A\right] \leq C\left(\frac{A}{t}\right)^m,$$

where $C := C(p, m, N)$ is a constant.

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 it 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 imagine our urn process isolated from the graph process and call adding a ball "a step" of the process. We use $s = 1, 2, \dots, D_t^*(S^*) - 2Nm$ to index the steps of the urn process.

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

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

And therefore if $A > 0$

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

Recalling that $r \ge m$ and $r + b = 2Nm$ and $\deg_t^*(x_i^*) = R_{D_t^*(S^*) - 2Nm}$ we get, using Lemma 6 with $t_0 = N$,

$$\begin{aligned}
\mathbf{Pr}\left[\deg_t^*(x_i^*) \le A\right] &\le \mathbf{Pr}\left[\deg_t^*(x_i^*) \le A \,\middle|\, D_t^*(S^*) \ge \frac{2pm}{1+p} t - t^{1/2} \ln t\right] \\
&\qquad + \mathbf{Pr}\left[D_t^*(S^*) < \frac{2pm}{1+p} t - t^{1/2} \ln t\right] \\
&\le \mathbf{Pr}\left[R_s \le A \,\middle|\, s \ge \frac{2pm}{1+p} t - t^{1/2} \ln t - 2Nm\right] + e^{-p(\ln t)^2/8m^3} \\
&\le \frac{2^{2Nm} A^r}{r\left(\frac{2pm}{1+p} t - t^{1/2} \ln t - 2Nm\right)^r} + e^{-p(\ln t)^2/8m^3} \\
&\le C\left(\frac{A}{t}\right)^r \le C\left(\frac{A}{t}\right)^m .
\end{aligned}$$

□

Lemma 8. *Let $s > N$ and let $t \geq s$.*

$$\mathbf{Pr}\left[\deg_t^*(x_s^*) \geq (t/s)^{q/2}(\ln t)^3\right] \leq \exp\left(m - \frac{(\ln t)^2}{4}\right)$$

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

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

$$X_{\tau+1} = x + \text{Binomial}\left(m, \frac{qx}{2m\tau}\right) \tag{14}$$

and so

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

for any $\lambda \leq 1$.

Thus

$$\mathbf{E}\left[e^{\lambda X_{\tau+1}}\right] \leq \mathbf{E}\left[\exp\left(X_\tau\lambda\left(1 + \frac{q(1+\lambda)}{2\tau}\right)\right)\right]. \tag{15}$$

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

$$\lambda_\tau \leq \Lambda = \min\left\{1, \frac{1}{\ln(t/s)}\right\} \text{ for } \tau = s, \ldots, t, \tag{16}$$

then (15) implies

$$\mathbf{E}\left[e^{\lambda_{\tau+1}X_{\tau+1}}\right] \leq \mathbf{E}\left[e^{\lambda_\tau X_\tau}\right] \qquad \text{for } \tau = s+1, \ldots, t-1.$$

Hence,

$$\mathbf{E}\left[e^{\lambda X_t}\right] = \mathbf{E}\left[e^{\lambda_t X_t}\right] \leq \mathbf{E}\left[e^{\lambda_s X_s}\right] = e^{m\lambda_s}.$$

We can write

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

then

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

the 2 bounds $e^{\gamma+1/2t}$ where γ is Euler's constant,

$$\leq 2e^{q/2}\lambda(t/s)^{q/2}$$

and therefore we can take $\lambda = \frac{\Lambda}{2e^{q/2}}(s/t)^{q/2}$ and get (16).

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

$$\begin{aligned}\mathbf{Pr}(X_t \geq u) &\leq e^{m\lambda_s - \lambda u} \\ &\leq \exp\left(\Lambda m - \frac{\Lambda(\ln t)^3}{4}\right) \\ &\leq \exp\left(m - \frac{(\ln t)^2}{4}\right).\end{aligned}$$

□

Lemma 9. *If $0 < A \ll t$ then,*

$$\mathbf{Pr}\left[z_t - Z_t \leq A\right] \leq C\left(\frac{A}{t}\right)^m,$$

where $C := C(p, m, N)$ is a constant.

Proof Let C' be the constant from Lemma 7. Then

$$\mathbf{Pr}\left[z_t^* < 2A\right] \leq C'\left(\frac{2A}{t}\right)^m \tag{17}$$

Also, from Lemma 8,

$$\mathbf{Pr}\left[Z_t^* \geq t^{q/2}(\ln t)^3\right] \leq t\exp\left(m - \frac{(\ln t)^2}{4}\right). \tag{18}$$

Using Lemma 1, and putting (17) and (18) together we get that if $A \geq t^{1/2}$ and t is sufficiently large

$$\begin{aligned}\mathbf{Pr}\left[z_t - Z_t \leq A\right] &\leq \mathbf{Pr}\left[z_t^* - Z_t^* \leq A\right] \\ &\leq \mathbf{Pr}\left[z_t^* < 2A\right] + \mathbf{Pr}\left[Z_t^* \geq A\right] \\ &\leq C'\left(\frac{2A}{t}\right)^m + t\exp\left(m - \frac{(\ln t)^2}{4}\right) \\ &\leq 4^m C'\left(\frac{A}{t}\right)^m.\end{aligned}$$

□

5 Proof of Theorem 1

Fix $i \leq N$. It follows from the (sub)-martingale convergence theorem (see [7]) and Lemma 4, that

$$L = \lim_{t\to\infty} \frac{\mathbf{E}\left[\deg_t(s_i)\right]}{t} \textit{ exists.}$$

We have to show that L is strictly positive and bounded away from zero. But $L \geq m/N$ follows immediately from Lemmas 1 and 3. This proves the first part of Theorem 1.

Our proof of the second part of Theorem 1 is a little more complicated, due to the fact that we want to estimate $\bar{d}_k(n)/n$ reasonably accurately (as opposed to using martinagale convergence as we did in Part (a)). Let $\gamma_t = Z_t - z_t$ and let t_0 be the first time that $\gamma_t \geq n^{q/2}$. Let $t_1 > t_0$ be the first time after t_0 such that $\gamma_t \leq m$.

Notice that in the time interval $[t_0, t_1]$, S_t is fixed and therefore, conditional on $G_{t_0} = G^*_{t_0}$, processes $\mathcal{P}$ and $\mathcal{P}^*$ coincide for every t such that $t_0 \leq t \leq t_1$.

Thus,

$$
\begin{aligned}
\mathbf{E}\left[d_k(G_n)\right] &= \mathbf{E}\left[d_k(G_n)|t_0 > n\right]\mathbf{Pr}\left[t_0 > n\right] + \mathbf{E}\left[d_k(G_n)|t_1 \leq n\right]\mathbf{Pr}\left[t_1 \leq n\right] \\
&\quad + \sum_{t=1}^{n}\sum_{G\in\mathcal{G}_t}\mathbf{E}\left[d_k^*(n)|G_{t_0} = G\right]\mathbf{Pr}\left[t_0 = t, G_t = G\right] \\
&= O\left(n\right)\left(\mathbf{Pr}\left[t_0 > n\right] + \mathbf{Pr}\left[t_1 \leq n\right]\right) \\
&\quad + \sum_{t=1}^{n}\sum_{G\in\mathcal{G}_t}\left(\frac{2n}{2+mq}\prod_{i=m+1}^{k}\frac{i-1}{i+2/q} + \tilde{O}\left(t + n^{q/2}\right)\right)\mathbf{Pr}\left[t_0 = t, G_t = G\right] \\
&= \frac{2n}{2+mq}\prod_{i=m+1}^{k}\frac{i-1}{i+2/q} + \\
&\quad O\left(n\right)\left(\mathbf{Pr}\left[t_0 > n\right] + \mathbf{Pr}\left[t_1 \leq n\right]\right) + \tilde{O}\left(n^{q/2}\right) + \tilde{O}\left(1\right)\mathbf{E}\left[t_0|t_0 \leq n\right] \qquad (19)
\end{aligned}
$$

It only remains to show that the contributions from (19) are $\tilde{O}\left(n^{q/2}\right)$. Let C be the constant defined in Lemma 9

Firstly,

$$
\mathbf{Pr}\left[t_0 > n\right] \leq \mathbf{Pr}\left[\gamma_n \leq n^{q/2}\right] \leq C\left(\frac{n^{q/2}}{n}\right)^m = o(n^{-1})
$$

and, as $t_0 \geq n^{q/2}/m$

$$
\mathbf{Pr}\left[t_1 \leq n\right] \leq \sum_{t=n^{q/2}/m}^{n}\mathbf{Pr}\left[\gamma_t \leq m\right] \leq \sum_{t=n^{q/2}/m}^{n} C\left(\frac{m}{t}\right)^m = O(n^{-mq/2}) = O(n^{-1}).
$$

Finally,

$$
\mathbf{E}\left[t_0|t_0 \leq n\right] \leq \sum_{t=N}^{n}\mathbf{Pr}\left[t_0 \geq t\right] \leq n^{q/2} + \sum_{t=n^{q/2}}^{n} C\left(\frac{n^{q/2}}{t}\right)^m = O\left(n^{q/2}\right).
$$

□

6 Concluding remarks

We have shown that modeling the influence of a search engine within the preferential attachment framework leads to a qualitative change in the familiar power-law degree distribution. Each of a collection of celebrities captures a constant fraction of the total degree of the graph, and the degree of the remaining nodes follow a steeper power law.

Our model differs from reality in many obvious ways: edges are undirected, outlinks are not modified after creation, pages do not die, there is no topic-based clustering, and there is no propensity toward forming bipartite cores as in the copying model.

Despite these limitations, our results lend support to recent articles by political scientists [10] in the popular press expressing apprehension about the extent to which search engines concentrate the collective attention of Web surfers to "mainstream" Web sites. Another study [13] involving live users rating jokes confirms the existence of the entrenchment effect, and shows that it can be reduced by limited randomization of the ranked list.

However, there is no verdict yet on the severity of the entrenchment effect of search engines in practice. A recent study [9] claims that the use of search engines actually has an egalitarian effect, in part owing to the diversity of query words used in searches. Enhancing entrenchment models with link-copying and query effects would be natural candidates for future work.

Acknowldegement: We would like to thank the referee for his/her persistence in suggesting the use of martingale convergence in the proof of Theorem 1(a).

References

[1] N. Alon and J. Spencer, The Probabilistic Method, Second Edition, Wiley Interscience, 2000.

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

[3] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW Conference, 1998. `http://www7.scu.edu.au/00/index.htm`.

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

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

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

[7] R. Durrett, Probability: Theory and examples, Wadsworth, Belmont California, 1991.

[8] A. Flaxman, A.M. Frieze and T.I. Fenner, High degree vertices and eigenvalues in the preferential attachment graph, Internet Mathematics 2 (2005) 1-20.

[9] S. Fortunato, A Flammini, F. Menczer, and A. Vespignani. The egalitarian effect of search engines. arXiv.org preprint at `http://arxiv.org/abs/cs/0511005`.

[10] 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.

[11] 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.

[12] 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.

[13] S. Pandey, S. Roy, C. Olston, J. Cho and S. Chakrabarti. Shuffling a stacked deck: The case for partially randomized ranking of search engine results. VLDB Conference, pages 781–792, 2005.

[14] D. M. Pennock, G. W. Flake, S. Lawrence, C. Lee Giles, E. J. Glover. Winners don't take all: Characterizing the competition for links on the Web. Proceedings of the National Academy of Sciences, 99(8), pages 5207–5211, 2002.

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