Evolution of Page Popularity under Random Web Graph Models

Rajeev Motwani [*] Ying Xu[†]

Abstract

The link structure of the Web can be viewed as a massive graph. The preferential attachment model and its variants are well-known random graph models that help explain the evolution of the web graph. However, those models assign more links to older pages without reference to the quality of web pages, which does not capture the real-world evolution of the web graph and renders the models inappropriate for studying the popularity evolution of new pages.

We extend the preferential attachment model with page quality, where the probability of a page getting new links depends not only on its current degree but also on its quality. We study the distribution of degrees among different quality values, and prove that under discrete quality distributions, the degree sequence still follows a power law distribution. Then we use the model to study the evolution of page popularity. We show that for pages with the same quality, the older pages are more popular; if a younger page is better than an older page, then eventually the younger-and-better page will become more popular. We also use the model to study a randomized ranking scheme proposed earlier [18] and show that it accelerates popularity evolution of new pages.

1 Introduction

The link structure of the Web may be viewed as a (directed) graph in which each vertex is a web page, and each edge is a hyperlink between two pages. The web graph has some interesting properties, such as power law degree sequence and small diameter. A number of stochastic models for the web graph have been proposed to better understand and predict the statistical properties of the Web.

A popular web graph model is the *preferential attachment* model [1]. We refer to this model and its variants as *random web graph models*. In this model, vertices and edges are dynamically added to the graph, such that the probability that an existing vertex gets a new link depends positively on its current degree. The resulting process generates graphs whose degree sequences follow a power law distribution, which is an observed property of the Web [13].

Recently, there has been some interest in studying the evolution of page popularity [9, 10]: when a page is newly created, very few people know about it; gradually it will become more popular in terms of more incoming links and higher user traffic. We are interested in studying how page popularity evolves over time under random graph models, in which a natural measure of popularity of a page is the degree of its corresponding vertex. However, we find those models lack one important element — the quality of pages. They treat all pages as being of the same quality which results in older pages generally having higher degrees. For example, we consider the basic *preferential attachment* model [1] which treats the graph as being undirected. The graph starts with a single vertex with an edge pointing to itself. At each time step, we add one new vertex and one edge connecting the new vertex to an existing vertex; the probability that the new edge links to vertex u is proportional to the current degree of u, i.e., at time $t+1$, the probability is $d_u(t)/2t$ where $d_u(t)$ is the degree of u at time t. Consider a page created at time t_0. Denote by $d(t)$ its degree at time $t \geq t_0$, then $E[d(t+1)|d(t)] = d(t) + d(t)/2t$. Taking expectation on both sides, we obtain the recurrence

$$E[d(t+1)] = E[d(t)] + \frac{E[d(t)]}{2t}.$$

Solving the recurrence under the initial condition $E[d(t_0)] = 1$, we have that

$$E[d(t)] \sim (\frac{t}{t_0})^{\frac{1}{2}}.$$

Therefore, older vertices always have higher expected degrees under preferential attachment; while in the real world, the popularity of a page not only depends on how old it is, but also how good. In this paper, we extend random graph models with page quality, similar to the *fitness model* in [2], and then study popularity evolution under the extended model.

[*]Stanford University. E-mail: rajeev@cs.stanford.edu. Supported in part by NSF Grant ITR-0331640, and grants from Media-X and SNRC.

[†]Stanford University. E-mail: xuying@cs.stanford.edu. Supported in part by Stanford Graduate Fellowship and NSF Grant ITR-0331640.

1.1 Our Model: Quality-based Preferential Attachment

We use a model similar to the *fitness model* in [2], extending random graph models with page quality. For simplicity, we extend the basic preferential attachment model and treat the graph as undirected.

The graph starts with one vertex and a self-loop. At each time step t, a new vertex is added and an edge connects it to some existing vertex. Each vertex u has a quality $q(u)$ that is chosen from some fixed probability distribution. The destination of the new edge is randomly chosen from all existing vertices in proportion to current degree weighted by quality, i.e., the probability of the new edge linking to u is $d_u(t)q(u)/\sum_v d_v(t)q(v)$ where $d_u(t)$ is the degree of u at time t.

The intuition of using $d \times q$ is as follows. The page quality q indicates the conditional probability that a user likes the page given that she has visited it. (To be interpreted as probability, q must be from $[0, 1]$, but it does not affect our results if all q's are multiplied by a common factor.) When a new page wants to link to some old page, it first randomly chooses to visit a page in the web with probability proportional to the popularity (current degree) of the pages; if it has visited page u, then with probability $q(u)$ it likes u and links to u, otherwise it repeats the procedure. Thus the probability that the new page links to u is proportional to the probability that it visits u multiplying the conditional probability that it likes u upon visiting, i.e., $d_u(t)q(u)$; after normalization, the probability equals to $d_u(t)q(u)/\sum_v d_v(t)q(v)$.

1.2 Our Results

We first study some basic properties of the quality-based preferential attachment model. We are mainly interested in two properties:

1. The quantity $DQ(t) = \sum_v d_v(t)q(v)$. This is the normalization factor that we will use to calculate linking probabilities. It also reflects the distribution of degrees among vertices of different qualities — $DQ(t)/2t$ is the average quality of the vertices (weighted by their degrees) in the graph and corresponds to the "quality-per-click" metric used in [18].

2. The overall degree sequence distribution. This can verify whether the model fits the real Web graph, whose degree sequence follows the power law distribution ([13]).

We analyze the model under discrete quality distribution, and prove the following results (see Theorems 2, 4 and 5 for a precise statement).

1. $E[DQ(t)] = Ct + o(t)$, where C is some constant.

2. The degree sequence follows a power law distribution.

The continuous quality distribution case has been studied in [2] and similar results hold. We summarize the results of [2] in Section 4.

Then we study the evolution of page popularity under this model. We define two parameters that measure popularity evolution. One is the *expected degree* $E[d_p(t)]$ which is the expected degree of page p at time t; and, the other is the *catch-up time*, defined as the time needed for a younger page with better quality to become more popular than an older page with poorer quality. We show that:

1. For a page p of quality q created at time t_0, denoted by $p(q, t_0)$, we have $E[d_p(t)] \sim (t/t_0)^{q/C}$, where C is some constant depending on the quality distribution.

2. Given two pages A(q_A, t_A) and B(q_B, t_B), where $t_A = \alpha t_B$, $q_A = \beta q_B$, $\alpha, \beta > 1$, the catch-up time $t = \alpha^{\frac{1}{\beta-1}} t_A$.

In other words, under this model, for pages with the same quality, the older pages are generally more popular; for pages with different qualities, eventually a younger page with better quality will be more popular than an older page with poorer quality, and the larger the quality gap, the shorter time it takes the younger page to catch up with the old page. This intuitively captures what happens in the real world. Note that the catch-up time does not depend on the quality distribution.

Finally, we study the impact of an alternative page ranking scheme on popularity evolution using the quality-based random graph model. The Partially Randomized Page Rank scheme was proposed by Pandey et al [18] to accelerate the popularity accumulation of new pages. Our analysis shows that introducing randomization in ranking does shorten the catch-up time, which measures a model's capability of popularizing new pages with good quality.

The rest of the paper is organized as follows. Section 2 reviews related work. In Section 3 and Section 4, we analyze the quality-based preferential attachment model under different quality distributions. Then we study the evolution of page popularity and the impact of randomized page ranking in Section 5. We conclude in Section 6 by pointing out some directions for further work.

2 Related Work

Barabasi and Albert [1] were among the first to propose a generative model for the web graph, called *preferential attachment*, to explain the power law distribution of its degree sequence. Kumar et al [14] proposed the *copying model*, in which a new page picks a random "prototype" page and with some probability copies outlinks from the prototype, and formally proved the power law degree distribution and some other properties of the resulting graph. The two models are similar in spirit: vertices and edges are dynamically added to the graph, and the probability that an existing vertex gets a new link depends positively on its current degree. In this class of models, many variations and extensions have been proposed [3, 5, 6, 7, 8, 11, 17].

Recently, Cho and Roy [9, 10] have studied how the popularity of web pages evolves over time and how search engines influence this process. Their results are intriguing, but the analysis lacks a rigorous model. They did not model the structure of the web graph, thus had to rely on the assumption that when users perform random walks on the web graph, the visit rate for a page (which is the stationary distribution of random walk [16, 4]) is proportional to the fraction of web users who like the page. They did not justify the assumption in their papers.

Our quality-based preferential attachment model is very similar to the fitness model of [2]. While [2] only gives heuristic analysis of continuous quality distribution, we also consider the discrete distribution and prove our results formally. What is more, we apply the model to study popularity evolution of webpages and evaluate alternative ranking schemes.

3 Discrete Quality Distribution

In this section, we study some basic properties of the quality-based preferential attachment model for discrete quality distributions. We start with a simple quality distribution where there are two possible quality values and a page can be either of the two with equal probabilities. We will compute $E[DQ(t)]$ and prove that $DQ(t)$ sharply concentrates around its expectation. We will also prove that the overall degree sequence follows a power law distribution. The theoretical analysis is confirmed by simulation results (Section 3.3). Finally, we will generalize the results to other discrete distributions (Section 3.4).

3.1 Computing $DQ(t)$

We consider the quality distribution where there are two possible quality values and a page can be either of the two with equal probabilities. Without loss of generality, we assume that the two quality values are $q_0 = 1$ and $q_1 = q > 1$.

We are interested in the quantity $DQ(t) = \sum_v d_v(t)q(v)$, which is the normalization factor of linking probabilities and also reflects the average quality of the vertices weighted by their degrees. Let $d(t)$ be the total degree that is associated with vertices with quality q_0 at time t, i.e. $d(t) = \sum_{u:q(u)=q_0} d_u(t)$. Once we know $d(t)$, we immediately get $DQ(t) = d(t) + [2t - d(t)]q$.

Given $d(t)$, the increment of d at the next time step consists of two parts: with probability $\frac{1}{2}$ the new vertex has quality q_0 and adds 1 degree to $d(t+1)$; the new edge may connect to old vertices with quality q_0. Therefore, the conditional expectation of $d(t+1)$ given $d(t)$ is

$$\begin{aligned} E[d(t+1)|d(t)] &= d(t) + \frac{1}{2} + \sum_{u:q(u)=q_0} \frac{d_u(t)q(u)}{DQ(t)} \\ &= d(t) + \frac{1}{2} + \frac{q_0 \sum_{u:q(u)=q_0} d_u(t)}{DQ(t)} \\ &= d(t) + \frac{1}{2} + \frac{d(t)}{DQ(t)} \end{aligned}$$

The last equation uses the definition of $d(t)$ and the assumption that $q_0 = 1$.

Taking expectation on both sides, we get

$$E[d(t+1)] = E[d(t)] + \frac{1}{2} + \frac{E[d(t)]}{DQ(t)} \quad (1)$$

Let $a(t)$ be the expectation of the fraction of $d(t)$ over the total degrees, i.e. $a(t) = E[d(t)/2t]$. Assume in the steady state $a(t)$ is a constant independent on t, i.e. $a(t) \equiv a$. Then Equation (1) becomes

$$2(t+1)a = 2ta + \frac{1}{2} + \frac{a}{a + (1-a)q}$$

Simplifying the equation, we get

$$4(q-1)a^2 - (5q-3)a + q = 0$$

We can solve a from the above equation, and in the steady state, $d(t) = 2ta$ and $DQ(t) = 2t[a + (1-a)q]$. Note that the left side of the above equation is always greater than 0 when $a = \frac{1}{4}$, and less than 0 when $a = \frac{1}{2}$, so there must exist a root in $(\frac{1}{4}, \frac{1}{2})$.

Now consider the degree sequence. First consider the subset of vertices with quality q_0. The distribution of degrees within the subset follows preferential attachment, so the degree subsequence follows the power law distribution. So does the degree subsequence of vertices with quality q_1. Merging two sequences with power law distributions, we still get a power law distributed sequence.

3.2 Formal Proof

We now formalize the above analysis.

The first theorem states that the expectation of $d(t)$ is $2t(a+o(1))$.

Theorem 1 *Let $d(t)$ be the sum of degree over vertices with quality q_0 at time t. Then*

$$2at - 2C_1 t^{\frac{3}{4}} \le E[d(t)] \le 2at + 2C_2 t^{\frac{3}{4}}$$

where C_1 and C_2 are two constants, a is a constant in $(\frac{1}{4}, \frac{1}{2})$ satisfying $4(q-1)a^2 - (5q-3)a + q = 0$.

Proof: We have the recurrence

$$E[d(t+1)] = E[d(t)] + \frac{1}{2} + \frac{E[d(t)]}{E[d(t)] + (2t - E[d(t)])q}$$

Prove the result by induction. First we prove the upper bound. We need to show

$$2at + 2C_2 t^{\frac{3}{4}} + \frac{1}{2} + \frac{2at + 2C_2 t^{\frac{3}{4}}}{2at + 2C_2 t^{\frac{3}{4}} + (2t - 2at - 2C_2 t^{\frac{3}{4}})q} \le 2a(t+1) + 2C_2(t+1)^{\frac{3}{4}}$$

or equivalently,

$$\frac{C_2 t^{-\frac{1}{4}}}{(C_2 t^{-\frac{1}{4}} + a + (1 - a - C_2 t^{-\frac{1}{4}})q)(a + (1-a)q)} \le 2C_2(t+1)^{\frac{3}{4}} - 2C_2 t^{\frac{3}{4}}$$

Because $t^{-\frac{1}{4}} = o(1)$ and $a \in (0, \frac{1}{2})$, the denominator of the left side is at least $(1-a^2)q \ge \frac{3}{4}q \ge \frac{3}{4}$. Thus we only need to show $\frac{2}{3}t^{-\frac{1}{4}} \le (t+1)^{\frac{3}{4}} - t^{\frac{3}{4}}$, which can be easily proved by expanding $(t+1)^{\frac{3}{4}}$ on the left side.

The lower bound can be proved similarly. □

The expectation of $DQ(t)$ is straightforward given $d(t)$, shown by the following corollary.

Corollary 2 $E[DQ(t)] = 2t(a + (1-a)q) + o(t)$.

The next theorem shows sharp concentration of $d(t)$, implying sharp concentration of $DQ(t)$.

Theorem 3

$$Pr[|d(t) - E[d(t)]| \ge \lambda] \le \ exp(-\frac{\lambda^2}{\tilde{O}(t^{\frac{9}{5}})}) + \tilde{O}(\frac{1}{t})$$

The notation $\tilde{O}$ means an additional polylog factor, so $\tilde{O}(t^{\frac{9}{5}}) = O(t^{\frac{9}{5}} \log^k t)$ for some constant k. This theorem shows $d(t)$ sharply concentrates around its expectation: let $\lambda = \tilde{O}(t^{\frac{9}{10}}) = o(t)$, the probability that $d(t)$ deviates from its expectation for more than λ is $\tilde{O}(\frac{1}{t}) = o(1)$. Note that $E[d(t)] = 2ta + o(t)$, so the deviation $\lambda = o(t)$ is very small compared to the expectation.

Proof: For fixed t, consider the sequence of $E[d(t)|d(k)]$, where $k = 0, 1, 2, \ldots, t$. This forms a martingale because for any k

$$E[E[d(t)|d(k)]] = E[d(t)],$$

i.e., the expectations of the variables in the sequences remain the same.

We use Azuma's inequality (see, for example, Section 4.4 in [15]) to prove the theorem.

(**Azuma's inequality**) Let $X_0, X_1, \ldots$ be a martingale such that for each k, $|X_k - X_{k-1}| \le c_k$, then

$$Pr[|X_t - X_0| \ge \lambda] \le 2exp(-\frac{\lambda^2}{2\sum_{k=1}^{t} c_k^2})$$

Note that in this martingale of $E[d(t)|d(k)]$,
$E[d(t)|d(0)] = E[d(t)]$, and
$E[d(t)|d(t)] = d(t)$.
Therefore to prove the theorem, we only need to find a sequence of c_k's satisfying the following two conditions:

$$|E[d(t)|d(k)] - E[d(t)|d(k-1)]| \le c_k \quad (2)$$

$$\sum_k c_k^2 = \tilde{O}(t^{\frac{9}{5}}) \quad (3)$$

$d(k)$ is non-decreasing on k and increases by at most 2 at each time step, i.e. $d(k) \le d(k+1) \le d(k) + 2$, so both $E[d(t)|d(k-1)]$ and $E[d(t)|d(k)]$ are at most $E[d(t)|d(k) = d(k-1) + 2]$ and at least $E[d(t)|d(k) = d(k-1)]$, and hence

$$|E[d(t)|d(k)] - E[d(t)|d(k-1)]| \le E[d(t)|d(k) = d(k-1)+2] - E[d(t)|d(k) = d(k-1)]$$

Therefore if we can bound

$$E[d(t)|d(k) = x+2] - E[d(t)|d(k) = x] \le c_k,$$

where x is any possible value of $d(k-1)$, then Inequality (2) is satisfied.

Let $e_k(i,x) = E[d(i)|d(k) = x]$; let $\Delta_k(i,x) = e_k(i, x+2) - e_k(i,x) = E[d(i)|d(k) = x+2] - E[d(i)|d(k) = x]$. The recurrence of Δ on i shows that starting from either $d(k) = x$ or $d(k) = x+2$, how the different of $d(i)$ evolves as the time i increases. We are interested in time $i = t$ and will find c_k such that $c_k \ge max_x \Delta_k(t,x)$.) The recurrence of Δ on i is as follows (for simplicity we omit the parameter x and use $\Delta_k(i)$ and $e_k(i)$).

$$\Delta_k(i+1) = \Delta_k(i) + \frac{e_k(i) + \Delta_k(i)}{e_k(i) + \Delta_k(i) + [2i - (e_k(i) + \Delta_k(i))]q}$$

$$
\begin{aligned}
& - \frac{e_k(i)}{e_k(i) + (2i - e_k(i))q} \\
= \quad & \Delta_k(i) + \frac{2iq\Delta_k(i)}{e_k(i) + \Delta_k(i) + [2i - (e_k(i) + \Delta_k(i))]q} \\
& * \frac{1}{e_k(i) + (2i - e_k(i))q} \\
\leq \quad & \Delta_k(i) + \frac{2iq\Delta_k(i)}{q(4i^2 - e_k(i)^2)} \\
= \quad & \Delta_k(i)(1 + \frac{2i}{4i^2 - e_k(i)^2})
\end{aligned}
$$

$\Delta_k(k) = 2$.

Note that if $e_k(i,x) \leq \frac{4i}{3}$, then $\Delta_k(i+1,x) \leq (1 + \frac{9}{10i})\Delta_k(i,x)$; solving the recurrence with the initial condition $\Delta_k(k,x) = 2$, we get $\Delta_k(i) \leq 2(\frac{i}{k})^{9/10}$ regardless of the value of x.

We consider two cases:

(1) $k \geq C \ln t$. Consider the probability that at time k more than $\frac{2}{3}$ fraction of the total degree is associated with vertices with quality q_0, i.e., $e_k(k,x) = d(k) \geq \frac{4k}{3}$. This probability is less than the probability that more than $\frac{2}{3}$ fraction of the vertices have quality q_0. Each vertex has quality q_0 with probability $\frac{1}{2}$ independently; according to Chernoff bound (see Section 4 in [15]), the probability that more than $\frac{2}{3}$ fraction of the vertices have quality q_0 is bounded by $e^{-O(k)}$. Choosing appropriate constant C, the probability is $O(t^{-2})$. Then using union bound over all time steps k from the beginning up to time t, $Pr[\exists k : d(k) \geq \frac{4k}{3}] = O(1/t)$.

We have proved with high probability, for all $k \geq C \ln t$ $d(k) < \frac{4k}{3}$. Now only consider the case that $d(k) < \frac{4k}{3}$, $\forall k \geq C \ln t$. Note that if $e_k(i_0, x) \leq \frac{4i_0}{3}$, then $e_k(i,x) \leq \frac{4i}{3}$ holds for all time $i > i_0$. Particularly, let $i_0 = k$, we have $e_k(i,x) \leq \frac{4i}{3}$ for all k and $i \geq k$. Therefore, $\Delta_k(i+1,x) \leq (1 + \frac{9}{10i})\Delta_k(i,x)$, and $\Delta_k(i,x) \leq 2(\frac{i}{k})^{9/10}$ for any x given that $d(k) < \frac{4k}{3}$, $\forall k \geq C \ln t$. Hence, when $k \geq C \ln t$, with probability $1 - O(1/t)$

$$
|E[d(t)|d(k)] - E[d(t)|d(k-1)]| = \Delta_k(t,x) \leq 2(\frac{t}{k})^{9/10}.
$$

Let $c_k = 2(\frac{t}{k})^{9/10}$, then Inequality (2) is satisfied. Summing up c_k^2 for all k in this case, $\sum_{k=C\ln t}^{t} c_k^2 = O(t^{9/5})$.

(2) $k < C \ln t$. We claim that even if $d(k) = 2k$, after $k' = C_1 k \ln t$ time steps, $d(k') \leq \frac{4k'}{3}$ with high probability. The reason is as follows. After $k \ln t$ steps, the expected degree associated with q_1 vertices is at least $k \ln t/2$, even if all new edges link to q_0 vertices; so at this time, $d(t)/2t < 3/4$. After another $2k \ln t$ steps, the expected degree of q_1 vertices is at least $k \ln t/2 + k \ln t + 2k \ln t/4$, because at least $1/4$ of the $2k \ln t$ new edges link to q_1 vertices. Now $d(t)/2t < 2/3$, i.e., at time $k' = 3k \ln t$, we expect $d(k') \leq \frac{4k'}{3}$. By setting the constant a bit larger, we can apply Chernoff bound to show that $d(k') \leq \frac{4k'}{3}$ with probability $1 - O(1/t)$. Using union bound, the probability that $d(k') \leq \frac{4k'}{3}$ holds for all $k < C \ln t$ is $1 - O(1/t) * C \ln t = 1 - \tilde{O}(1/t)$.

Therefore starting from time $k' = C_1 k \ln t$, we have $d(k') \leq \frac{4k'}{3}$ with high probability. Similar as case (1), $\Delta_k(i,x) \leq \Delta_k(k',x)(\frac{i}{k'})^{9/10}$. Since $\Delta_k(k',x)$ is at most $2k'$, we have

$$
\Delta_k(t,x) \leq 2C_1 k \ln t (\frac{t}{C_1 k \ln t})^{9/10}
$$

for any x. Let $c_k = 2C_1 k \ln t(\frac{t}{C_1 k \ln t})^{9/10}$. Summing up c_k^2 for case (2), $\sum_{k=1}^{C \ln t} c_k^2 = O(t^{9/5})$.

Combining the two cases, with probability $1 - \tilde{O}(1/t)$, $\sum_{k=1}^{t} c_k^2 = O(t^{9/5})$. Applying Azuma's inequality, the theorem is proven. □

Finally we analyze the degree sequence.

Theorem 4 *The degree sequence follows a power law distribution with parameter* $\gamma = \frac{1}{3-4a} + 2$.

Note that $a \in (\frac{1}{4}, \frac{1}{2})$, so the power law parameter $\gamma \in (2.5, 3)$. Specially, when $q = 1$, it degenerates to the classic preferential model, and we get $a = \frac{1}{2}$, hence $\gamma = 3$. On the other hand, when $q \to \infty$, $a \to \frac{1}{4}$, $\gamma \to 2.5$.

Proof: Let $X0_i(t)$ be the number of vertices with quality q_0 and degree i at time t; $X1_i(t)$ be the number of vertices with quality q_1 and degree i; $X_i(t) = X0_i(t) + X1_i(t)$ be the total number of vertices with degree i.

First consider only those vertices with quality q_0. At time t, the number of q_0 vertices is $t + o(t)$ and the total degree is $2at + o(t)$ with high probability. The degree distribution among those vertices is similar to the classic preferential attachment model. Consider the number of vertices with degree i at time $t+1$ ($X0_i(t+1)$). Those vertices have degrees either i or $i-1$ at time t. A vertex with degree $i-1$ now has degree i only if it gets the link from the new vertex. Conditioned on that the new edge links to a q_0 vertex, the probability that a particular q_0 vertex with degree $i-1$ at time t has degree i at time $t+1$ is $(i-1)/(2at+o(t))$, and the probability that some q_0 vertex with degree $i-1$ has degree i at time $t+1$ is $(i-1)X0_{i-1}(t)/(2at+o(t))$. On the other hand, vertices with degree i at time t may get linked to the new edge and no longer have degree i; conditioned on that the new edge links to a q_0 vertex, the probability that some vertex in $X0_i(t)$ is not in $X0_i(t+1)$ is $iX0_i(t)/(2at + o(t))$. Finally, the probability that the new edge links to a q_0 vertex is

$$\begin{aligned}\frac{2at+o(t)}{2t(a+(1-a)q)+o(t)} &= \frac{a}{a+(1-a)q}+o(1)\\ &= 2a-\frac{1}{2}+o(1)\end{aligned}$$

The last equation holds because $4(q-1)a^2-(5q-3)a+q=0$. Therefore the growth rate of $X0_i$ can be expressed as

$$\frac{dX0_i(t)}{dt} = \frac{(i-1)X0_{i-1}(t)-iX0_i(t)}{2at+o(t)} * (2a-\frac{1}{2}+o(1))$$

Assume in the steady state $X0_i(t)=c_it$ for some constant c_i, then for large i,

$$\begin{aligned}\frac{c_i}{c_{i-1}} &\sim \frac{(2a-\frac{1}{2})(i-1)}{2a+(2a-\frac{1}{2})i}\\ &= 1-\frac{4a-\frac{1}{2}}{2a+(2a-\frac{1}{2})i}\\ &\sim 1-\frac{8a-1}{4a-1}*\frac{1}{i}\\ &\sim (\frac{i}{i-1})^{-\frac{1}{4a-1}-2}\end{aligned}$$

For above to hold, we must have

$$X0_i(t) \sim ti^{-\frac{1}{4a-1}-2}$$

The above analysis is standard in preferential attachment type models, and can be made more precise along the lines of proof in [14, 6].

Similarly, the degree sequence of vertices with quality q_1 follows a power law distribution:

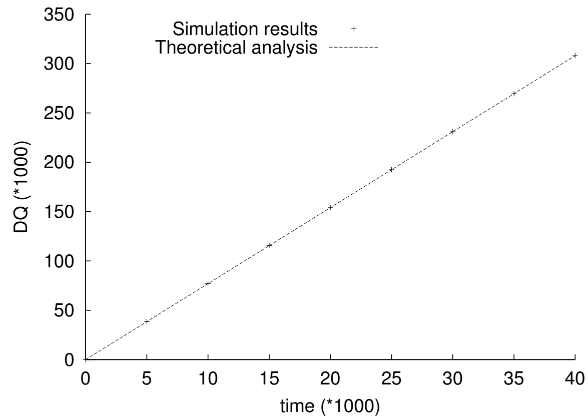
$$X1_i(t) \sim ti^{-\frac{1}{3-4a}-2}$$

Consider the overall degree sequence $X_i(t)=X0_i(t)+X1_i(t)$. For large i, $X1_i(t)$ dominates $X0_i(t)$, so $X_i(t)$ also follows a power law distribution with parameter $\frac{1}{3-4a}+2$. □

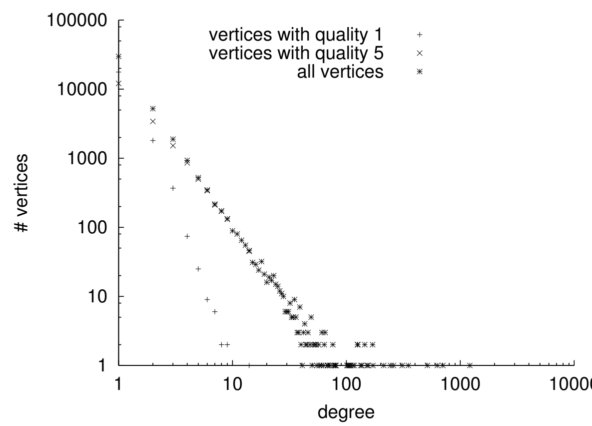
3.3 Experimental Results

We generate random graphs according to our quality-based preferential attachment model, and measure $DQ(t)$ and degree distribution of the result graphs. The simulation results confirm our analysis.

Figure 1 shows the simulation results when quality of a page is either 1 or 5, with equal probability. Figure 1(a) shows the evolution of $DQ(t)$ (Y axis) over time (X axis): the line gives $DQ(t)$ values predicted by the analysis (solving the equation, we get $a \approx 0.287$ and $c \approx 3.85$); the points are measured values from the simulation. We can see the experimental result fits our analysis very well. Figure 1(b) is the Log-Log plot of degree distribution. We plot the sequences $X0_i(t)$, $X1_i(t)$ and $X_i(t)$ when $t=40000$, where $X0_i(t)$ is the number of vertices with quality 1 and degree i at time t, $X1_i(t)$ is the number of vertices with quality 5 and degree i at time t, and $X_i(t)$ the total number of vertices with degree i at time t. Each of the three sequences forms a line in the Log-Log plot, indicating a power law distribution. When degree i is large, $X1_i$ dominates $X0_i$, and X_i is close to $X1_i$.



(a) DQ(t)



(b) degree sequence

Figure 1: **Simulation results when the quality of a page is either 1 or 5 with equal probability.** Figure (a) shows the values of $DQ(t)$ at different times; Figure (b) shows the degree distribution at $t=40000$, i.e. when there are 40000 vertices in the graph.

3.4 Generalization to other Discrete Distributions

We can generalize the quality distribution to any discrete distribution. Suppose the possible qualities are $q_1, \ldots, q_k$, with probabilities $p_1, \ldots, p_k$ respectively. Let $d_i(t)$ be the sum of degrees of vertices with quality q_i. Then similar to Equation (1) in Section 3.1, we have

$$E[d_i(t+1)] = E[d_i(t)] + p_i + \frac{E[d_i(t)]q_i}{DQ(t)}$$

Again assume in the steady state $E[d_i(t)] = 2ta_i$ for some constant a_i. Then the equations are:

$$2a_i = p_i + \frac{a_i q_i}{\sum_{j=1}^{k} a_j q_j}, i = 1 \ldots k$$

And $E[DQ(t)] = 2t \sum_{j=1}^{k} a_j q_j$.

4 Continuous Quality Distribution

For completeness, we include briefly in this paper the analysis from [2] on the continuous quality distribution.

Let $f(q)$ be the density function of the quality distribution, defined on the range $[a, b]$. Let $d_q(t)$ be the total degree of vertices whose quality is in the range $[q, q+dq)$. Then

$$E[d_q(t+1)] = E[d_q(t)] + f(q)dq + \frac{E[d_q(t)]q}{DQ(t)}$$

Assume in the steady state $E[d_q(t)] = 2ta_q$ for some constant a_i and $DQ(t) = 2tc$. Then $a(q)$ satisfies the following formula:

$$2(t+1)a(q) = 2ta(q) + f(q)dq + \frac{a(q)q}{c}$$

So we get

$$a(q) = \frac{cf(q)dq}{2c - q}$$

On the other hand $c = \int_a^b a(q)qdq$. Combining the two equations, we get

$$1 = \int_a^b \frac{qf(q)}{2c-q} dq$$

As for the degree distribution, the probability that a vertex has degree i is given by the integral

$$\int_a^b f(q) \frac{2c}{q} i^{-\frac{2c}{q} - 1} dq$$

The results are summarized in the following theorem.

Theorem 5 *[2] When the qualities follows a continuous distribution with density function $f(q)$,*

(1) $E[DQ(t)] = 2tc + o(t)$, where c is a constant satisfying $1 = \int_a^b \frac{qf(q)}{2c-q} dq$.

(2) The degree sequence is $X_i(t) \sim t \int_a^b f(q) \frac{2c}{q} i^{-\frac{2c}{q}-1} dq$.

Particularly, if the qualities follow the uniform distribution on $[0, 1]$, then c is the solution of $c \ln \frac{2c}{2c-1} = 1$, thus $c \approx 0.627$, i.e. $E[DQ(t)] \approx 1.255t$. The degree sequence is $X_i(t) \sim ti^{-2c-1} \ln i$, which is very close to the power law distribution with parameter 2.255. Note that in the continuous case, not all quality distributions render power law distributed degree sequences.

5 Evolution of Page Popularity

Having set up the model, we now examine the evolution of page popularity under the model. A natural indicator of page popularity under random graph models is its degree. We define two measures.

(1) *expected degree*, $E[d_p(t)]$: the expected degree of page p at time t;

(2) *catch-up time*: the time it takes a younger page A with better quality to become more popular than an older page B with poorer quality, i.e., the smallest time t satisfying $E[d_A(t)] \geq E[d_B(t)]$.

We compute expected degree and catch-up time in the quality-based preferential attachment model.

Denote by $p(q, t_0)$ a page created at time t_0 and having quality q. Denote by $d_p(t)$, or abbreviated as $d(t)$ if p is clear from the context, the degree of page p at time $t \geq t_0$. We have the recurrence

$$E[d(t+1)|d(t)] = d(t) + \frac{d(t)q}{DQ(t)}$$

with the initial condition $d(t_0) = 1$. From Corollary 2 and Theorem 5, $DQ(t) = Ct + o(t)$ with high probability for both discrete and continuous quality distribution, where C is some constant depending on the quality distribution. We omit the $o(t)$ term and assume $DQ(t) = Ct$. Taking expectation on both sides, the above recurrence becomes

$$E[d(t+1)] = E[d(t)] + \frac{E[d(t)]q}{Ct}; E[d(t_0)] = 1$$

Solving the recurrence, we get

$$\begin{aligned} &E[d(t)] \\ &= (1 + \frac{q}{C(t-1)})E[d(t-1)] \\ &= (1 + \frac{q}{C(t-1)})(1 + \frac{q}{C(t-2)}) \ldots (1 + \frac{q}{Ct_0})E[d(t_0)] \\ &\sim (\frac{t}{t_0})^{q/C} \end{aligned}$$

Since the degree of a vertex is a measure of the popularity of the page, we observe from the formula of $E[d_p(t)]$ that under quality-based preferential attachment, the popularity of a page positively depends on its quality and its age. Among pages with the same quality, older pages are generally more popular; if a younger page is better than an older page, then the older page remains its advantage for a certain period of time, but eventually the younger (better) page will be more popular. This is intuitively what happens in the real world.

Next we compute the catch-up time. Consider two pages: A created at time t_A with quality q_A and B created at time t_B with quality q_B. A is created later but with better quality. Assume $t_A = \alpha t_B$, $q_A = \beta q_B$, where α and β are constants greater than 1. We want to compute the time t when $E[d_A(t)] = E[d_B(t)]$. Plugging the expression for $E[d(t)]$, the equation becomes

$$(\frac{t}{t_A})^{q_A/c} = (\frac{t}{t_B})^{q_B/c}$$

or equivalently,

$$q_A \ln \frac{t}{t_A} = q_B \ln \frac{t}{t_B}$$

The solution to the equation does not depend on C:

$$t = \alpha^{\frac{1}{\beta-1}} t_A$$

For example, if $t_A = 10t_B$, $q_A = 2q_B$, then it takes $t = 10t_A$ for page A to catch up with B. If $q_A = 5q_B$, then $t = 10^{0.25}t_A \approx 1.8t_A$, so it takes much shorter time for A to become more popular than B. The larger the quality gap, the shorter the catch-up time is.

We summarize the results in the following theorem. We assume $DQ(t) = Ct$, which is a reasonable approximation given Corollary 2 and Theorem 5.

Theorem 6 *Under quality-based preferential attachment model, assuming $DQ(t) = Ct$,*

(1) For a page $p(q, t_0)$, $E[d_p(t)] \sim (t/t_0)^{q/C}$, where C is some constant depending on the quality distribution;

(2) Given two pages $A(q_A, t_A)$ and $B(q_B, t_B)$, where $t_A = \alpha t_B$, $q_A = \beta q_B$, $\alpha, \beta > 1$, then catch-up time $t = \alpha^{\frac{1}{\beta-1}} t_A$.

A nice property about catch-up time is that it does not depend on the quality distribution. Therefore even though we do not know the true distribution of page qualities in the Web, we can still compute the catch-up time, assuming the Web evolves as the model suggests.

5.1 Impact of Partially Randomized Ranking Scheme on Page Popularity

Ideally we would like search engines to rank the web pages by their qualities, but current PageRank scheme ([16, 4]) ranks pages by their popularity, thus bias towards older pages. Several recent studies [8, 9, 12] worry that search engines inhibit new, high-quality pages from gaining popularity. To remedy this problem, a number of alternative page ranking schemes have been proposed to boost up new pages ([10, 18]), but it is very expensive to conduct large scale experiments to evaluate the effectiveness of new ranking schemes. Now with our quality-based random graph model, it is possible to formally evaluate and compare alternative ranking schemes, using *catch-up time* as the measure of a model's capability of popularizing new pages with good quality.

We take the *partially randomized ranking scheme* [18] as an example and analyze its impact on page popularity under random graph models. The *partially randomized ranking* adds randomness to PageRank result to boost up new pages: for each entry in the final ranking, with probability α, it returns a random page with low popularity; with probability $1-\alpha$, it returns the page with the highest PageRank value among those not yet appeared in the list.

We can model this ranking scheme as the following *randomized ranking model*: at each time step a new vertex is added with an edge linking it to an existing vertex; with probability α, the new edge links to one of the recent N vertices with probability proportional to their qualities, where N is some constant; with probability $1 - \alpha$, the new edge randomly links to any existing vertex proportionally to its degree multiplying quality.

The intuition behind the randomized ranking model is that the new page first randomly chooses a page to visit according to the randomized ranking scheme (with probability $1 - \alpha$ choose by the ranking, with probability α choose one recently created page uniformly at random); if it has visited some page, it likes the page with probability proportional to the quality of that page. Note that our graph model is a simplified model for the web graph by treating the web as undirected. PageRank is the stationary distribution of the random walk on the graph and degenerates to ranking by degrees on undirected graphs.

Before studying the catch-up time in the new model, we first compute the quantity $DQ(t) = \sum_v d_v(t)q(v)$.

For discrete quality distribution, suppose the possible qualities are $q_1, \ldots, q_k$, with probabilities $p_1, \ldots, p_k$ respectively. Let $d_i(t)$ be the sum of degrees of vertices with quality q_i at time t. From time step t to $t + 1$, the increment of d_i consists of three parts: the new vertex has quality q_i with probability p_i; with probability $1 - \alpha$, the new edge links to a q_i vertex with probability $d_i(t)q_i/DQ(t)$, as the analysis in Section 3.4; with probability α, the new edge links to a recent vertex according to its quality and the probability of this vertex having quality q_i is $q_ip_i/\sum_j q_jp_j$. To sum up,

$$E[d_i(t+1)] = E[d_i(t)] + p_i + \frac{E[d_i(t)]q_i}{DQ(t)}(1-\alpha) + \frac{q_ip_i}{E[q]}\alpha$$

where $E[q]$ is the mean of quality ($E[q] = \sum_j q_jp_j$).

Assuming in the steady state $E[d_i(t)] = 2ta_i$, the a_i's satisfy the following equations:

$$2a_i = p_i + \frac{a_i q_i(1-\alpha)}{\sum_{j=1}^{k} a_j q_j} + \frac{\alpha q_i p_i}{E[q]}, i = 1 \ldots k$$

And $E[DQ(t)] = 2t \sum_{j=1}^{k} a_j q_j$.

For the continuous distribution with density function $f(q)$ defined on the range $[a,b]$, let $d_q(t)$ be the total degree of vertices whose quality is in the range $[q, q+dq)$. Then

$$E[d_q(t+1)] = E[d_q(t)] + f(q)dq + \frac{E[d_q(t)]q}{DQ(t)}(1-\alpha) + \frac{qf(q)dq}{E[q]}\alpha$$

Assuming $E[d_q(t)] = 2ta_q$ and $DQ(t) = 2ct$, $a(q)$ satisfies the following formula:

$$2(t+1)a(q) = 2ta(q) + f(q)dq + \frac{a(q)q(1-\alpha)}{c} + \frac{\alpha q f(q)dq}{E[q]}$$

or

$$a(q) = \frac{cf(q)dq(1+\frac{\alpha q}{E[q]})}{2c - q(1-\alpha)}$$

Combining with the equation $c = \int_a^b a(q)qdq$, we get

$$1 = \int_a^b \frac{qf(q)(1+\frac{\alpha q}{E[q]})}{2c - q(1-\alpha)}dq$$

$DQ(t)/2t$ gives the average quality of the pages weighted by their degrees. This is also an interesting indicator of the goodness of the model because it corresponds to the "quality-per-click" metric used in [18]. *quality-per-click* measures the average quality of pages viewed by users; in the random graph model we assume that the probability of a page being visited is proportional to its degree, so quality-per-click maps to $DQ(t)$. In the randomized ranking model, given the quality distribution, theoretically we can compute DQ for any α, and find α maximizing DQ. However in reality it is hard to know the quality distribution, so we prefer the catch-up time metric which is independent on the quality distribution.

Now we compute the *expected degree* and *catch-up time* under the randomized ranking model. Let $E[q]$ be the mean of quality of the given quality distribution. The expected degree of page $p(q, t_0)$ is given by

$$E[d(t)] = \begin{cases} E[d(t-1)] + \frac{E[d(t-1)]q(1-\alpha)}{DQ(t-1)}, \\ \qquad \text{if } t > t_0 + N \\ E[d(t-1)] + \frac{E[d(t-1)]q(1-\alpha)}{DQ(t-1)} + \frac{q\alpha}{NE[q]}, \\ \qquad \text{if } t_0 < t \leq t_0 + N \\ 1, \qquad \text{if } t = t_0 \end{cases}$$

N is a constant, so for large $t_0 \gg N$, during time interval $[t_0+1, t_0+N]$,

$$\frac{q\alpha}{NE[q]} = \Theta(1) \gg \frac{d(t)q(1-\alpha)}{DQ(t)} = \Theta(\frac{1}{t})$$

Therefore $E[d(t_0+N)] \approx 1 + \frac{\alpha q}{E[q]}$. Assume $DQ(t) = C_1 t$, and we get

$$\begin{aligned} & E[d(t)] \\ = \;& (1 + \frac{q(1-\alpha)}{C_1(t-1)})E[d(t-1)] \\ = \;& (1 + \frac{q(1-\alpha)}{C_1(t-1)}) \ldots (1 + \frac{q(1-\alpha)}{C_1(t_0+N)})E[d(t_0+N)] \\ \sim \;& (\frac{t}{t_0+N})^{q(1-\alpha)/C_1}(1 + \frac{\alpha q}{E[q]}) \end{aligned}$$

Given two pages $A(q_A, t_A)$ and $B(q_B, t_B)$ with $q_A = \alpha q_B, t_A = \beta t_B$ $(\alpha, \beta > 1)$, we can compute the catch-up time t by solving the equation $E[d_A(t)] = E[d_B(t)]$:

$$t = (\frac{E[q] + \alpha q_B}{E[q] + \alpha q_A})^{\frac{C_1}{(1-\alpha)(q_A - q_B)}} \frac{(t_A + N)^{\frac{q_A}{(q_A - q_B)}}}{(t_B + N)^{\frac{q_B}{(q_A - q_B)}}} \qquad (4)$$

The above t is smaller than catch-up time in Theorem 6 regardless of the quality distribution, because the first term is always less than 1 given $q_A > q_B$, while the second term is close to $\alpha^{\frac{1}{\beta-1}} t_A$ when $t_A > t_B \gg N$. This indicates that introducing randomness in ranking does accelerate the popularity accumulation of new pages.

Theorem 7 *In randomized ranking model, the expected degree*

$$E[d(t)] \sim (\frac{t}{t_0+N})^{q(1-\alpha)/C_1}(1 + \frac{\alpha q}{E[q]});$$

the catch-up time is given by Equation (4), and is smaller than the catch-up time in the quality-based preferential attachment.

6 Conclusions and Future Work

In this paper, we extend the random graph model with page qualities, and analyze the quality-based preferential attachment model under discrete quality distributions. We study the distribution of degrees among different quality values, and prove that the degree sequence still follows a power law distribution. Then we study the evolution of page popularity under this model. We also use the model to analyze a randomized ranking scheme and show that introducing randomness in ranking accelerates popularity evolution of new pages.

An interesting research topic is to study the impact of search engines and different ranking schemes on popularity evolution using our quality-based random graph models. Several recent papers [8, 9, 12] worry that search engines inhibit new, high-quality pages from gaining popularity, and a number of new page ranking schemes have been proposed to remedy this problem ([10, 18]). It is interesting to formally evaluate and compare alternative ranking schemes and the impact of search engines on popularity evolution using our quality-based random graph models. We also hope the model will provide insights for designing a new ranking scheme that can order the web pages by their true quality.

In this paper, we introduce quality to the basic preferential attachment model. The model has some limitations such as treating the graph as undirected, allowing no deletion of pages and links, and allowing no links from older pages to new pages. A number of extensions to the basic preferential attachment model have been proposed to address those limitations (for example [3, 5, 6, 7]), and it will be interesting to incorporate quality in those extended models too. Since we want to use random graph models to study popularity evolution under different ranking schemes and in reality the correlation between PageRank and degree is relatively low [17], it will be of special interest to model PageRank in random graph models.

7 Acknowledgement

We would like to thank Zoltan Gyongyi for pointing out some related work; Junfeng Yang and Mayank Bawa for helpful comments.

References

[1] A-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science,* 286(15), pp. 509-512, 1999.

[2] G.Bianconi and A-L. Barabasi. Competition and Multiscaling in evolving Networks. *Europhysics Letters*, 54(4), pp. 436-442, 2001.

[3] B. Bollobas, C. Borgs, T. Chayes and O. Riordan. Directed scale-free graphs. *Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2003.

[4] S. Brin, R. Motwani, L. Page, and T. Winograd. What can you do with a Web in your Pocket? *Data Engineering Bulletin,* 21: 37-47, 1998.

[5] B. Bollobas and O. Riordan. Robustness and vulnerability of scale-free random graphs. *Internet Mathematics,* 1: 1-35, 2003.

[6] C. Cooper and A. Frieze. A general model of web graphs. *Random Structures and Algorithms*, 22(3), pp. 311-335, 2003.

[7] C. Cooper, A. Frieze and J. Vera. Random deletion in a scale-free random graph process. *Internet Mathematics,* 1: 463-483, 2003.

[8] S. Chakrabarti, Alan Frieze and Juan Vera. The influence of search engines on preferential attachment. *Proceedings of the 16th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2005.

[9] J. Cho and S. Roy. Impact of Web search engines on page popularity. *Proceedings of the Thirteenth International WWW Conference*, 2004.

[10] J. Cho, S. Roy and R.E. Adams. Page quality: In search of an unbiased web ranking. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2005.

[11] E. Drinea, M. Enachescu and M. Mitzenmacher. Variations on random graph models of the Web. *Harvard Computer Science Technical Report* TR-06-01.

[12] E. Drinea, A. Frieze and M. Mitzenmacher. Balls and Bins Models with Feedback. *Proceedings of the 13th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2002.

[13] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan and A.S. Tomkins. The Web as a Graph: measurements, models and methods. *Proceedings of the 5th Annual International Conference on Computing and Combinatorics*, 1999.

[14] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal. Stochastic Models for the Web Graph. *Proceedings of 41st IEEE Symposium on Foundations of Computer Science (FOCS)*, 2000.

[15] R. Motwani and P. Raghavan. *Randomized Algorithm.* Cambridge University Press, 1995.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report, Stanford Digital Libraries,* 1998.

[17] G. Pandurangan, P. Raghavan and E. Upfal. Using PageRank to Characterize Web Structure. *Proceedings of the 8th Annual International Conference on Computing and Combinatorics*, 2002.

[18] 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. *Proceedings of the 31st International Conference on Very Large Data Base (VLDB)*, 2005.