

# Impact of Search Engines on Page Popularity

Junghoo Cho  
UCLA Computer Science Department  
Los Angeles, CA 90095  
cho@cs.ucla.edu

Sourashis Roy  
UCLA Computer Science Department  
Los Angeles, CA 90095  
roys@cs.ucla.edu

## ABSTRACT

Recent studies show that a majority of Web page accesses are referred by search engines. In this paper we study the widespread use of Web search engines and its impact on the ecology of the Web. In particular, we study how much impact search engines have on the popularity evolution of Web pages. For example, given that search engines return currently “popular” pages at the top of search results, are we somehow penalizing newly created pages that are not very well known yet? Are popular pages getting even more popular and new pages completely ignored? We first show that this unfortunate trend indeed exists on the Web through an experimental study based on real Web data. We then analytically estimate how much longer it takes for a new page to attract a large number of Web users when search engines return only popular pages at the top of search results. Our result shows that search engines can have an immensely worrisome impact on the discovery of new Web pages.

## 1. INTRODUCTION

Since the arrival of the Web in early 90’s, the Web search engines have become an indispensable tool in our everyday life. When we seek information, we often go to our favorite search engine and look at the returned pages. Given the sheer quantity of information available on the Web, the widespread use of search engines is not surprising. An individual simply cannot read billions of pages available on the Web, so he gets help from search engines to zoom in to a small number of pages worth looking at.

Despite search engines’ usefulness, we note that their widespread use may introduce a significant bias to people’s perception of the Web. For example, in a recent news article, a Web commentator stated that “if your page is not indexed by Google, your page does not exist on the Web [18].” While this statement may be an exaggeration, it contains an alarming bit of truth. To find a page on the Web, many Web users go to Google (or their favorite search engine) issue keyword queries, and look at the results. If the users cannot find relevant pages after several iterations of keyword queries, they are likely to give up and stop looking for further pages on the Web. Therefore, a page that is not indexed by Google (or ranked at the bottom) is unlikely to be viewed by many Web users.

The main question that we may ask is, then, how search engines rank Web pages given a query. If search engines

*fairly* judge the “quality” and “relevance” of every page and return the pages of highest quality, this “search-engine bias” may not be a significant problem. Unfortunately, the quality of a page is a very subjective notion and difficult to measure in practice, so most existing search engines use a “link-popularity” metric, called PageRank, to measure the “quality” of a page [21]. Roughly speaking, the PageRank metric considers a page “important” or of “high quality” if the page is linked to by many other pages on the Web.<sup>1</sup> For example, Google puts a page at the top of a search result (out of all the pages that contain the keywords that the user issued) when the page is linked to by the most other pages on the Web [5].<sup>2</sup> In short, “currently popular” pages are repeatedly returned at the top of the search results by major search engines.

The problem of this popularity-based ranking is that it is inherently biased against unknown pages. That is, when search engines constantly return popular pages at the top of their search results, more Web users will “discover” and look at those pages, increasing their popularity even further. In contrast, a currently-unpopular page will not be returned by search engines (or ranked at the bottom), so few new users will discover those pages and create a link to it, pushing the page’s ranking even further down. This “rich-get-richer” phenomenon can be particularly problematic for the “high-quality” pages that were recently created. Even if a page is of high quality, the page may be completely ignored by Web users simply because its current popularity is very low. This situation is clearly unfortunate both for Web page authors and the overall Web users. New and valuable pages are ignored just because they have not been given a chance to be noticed by people.

In this paper, we investigate the magnitude of this search-engine bias through experimental and theoretical studies:

**1. Experimental study** (Section 3): We first study whether the “rich-get-richer” phenomenon is happening in the current Web by examining real Web data collected over 7 months. The result strongly indicates that this phenomenon is indeed happening. From our experimental data, we could observe that the top 20% of the pages with the highest number of incoming links obtained 70% of the new links after 7 months, while the bottom 60% of the pages obtained virtually no new incoming links during that period.

<sup>1</sup>More precise description of the PageRank metric is provided in Section 2.

<sup>2</sup>While Google takes more than 100 factors into account in determining the final ranking of a page [12], the core of their ranking algorithm is based on the PageRank metric.

**2. Theoretical study** (Sections 4 and 5): We then study how much bias search engines can potentially introduce to the popularity of Web pages by analyzing two theoretical models on how users discover new Web pages.

1. **Random-surfer model** (Section 4): We assume that Web users discover new pages purely by surfing randomly on the Web, just following links. They never use a search engine that recommends pages based on their current popularity. This model, thus, roughly captures the case when users are *not* influenced by search engine ranking.
2. **Search-dominant model** (Section 5): We assume that users always start exploring the Web by going to a search engine and looking at the result. Search engines are the only way for the users to discover new pages. This model, therefore, represents the case when the search engines dominate the browsing pattern of users.

By comparing the popularity evolution of Web pages under these two models, we will be able to tell how much bias search engines introduce to the popularity evolution of Web pages. Our result shows that search engines can have an immensely worrisome impact on new Web pages. Our model predicts that it takes 60 times longer for a new page to become popular under the search-dominant model than under the random-surfer model. That is, if it took one year for a page to become popular without search engines, it may take more than 60 years for the same page to become popular when search engines are heavily used!

In Section 2, we first provide a brief introduction to the PageRank metric, the primary ranking metric used by Google, to help the reader understand how search engines measure the importance of a page. Then in Section 3 we present the result from our Web experiments and examine how the popularity of Web pages evolves over time. Finally in Sections 4 and 5, we analyze the two theoretical models to investigate how much an impact search engines have on the popularity of Web pages.

## 2. PAGERANK AND POPULARITY

In this section, we explain the basic intuition of PageRank and how it is related to a random-surfer model. A reader familiar with PageRank may skip this section.

Intuitively, PageRank is based on the idea that a link from page  $p_1$  to  $p_2$  may indicate that the author of  $p_1$  is interested in page  $p_2$ . Thus, if a page has many links from other pages, we may conclude that many people are interested in the page and that the page should be considered “important” or “of high quality.” Furthermore, we expect that a link from an important page (say, the Yahoo home page) carries more significance than a link from a random Web page (say, some individual’s home page). Many of the “important” or “popular” pages go through a more rigorous editing process than a random page, so it would make sense to value the link from an important page more highly.

The PageRank metric  $PR(p)$ , thus, defines the importance of page  $p$  to be the sum of the importance of the pages that point to  $p$ . Thus, if many important pages point to  $p$ ,  $PR(p)$  will be high. More formally, consider page  $p_i$  that is pointed at by pages  $p_1, \dots, p_m$ . Let  $c_j$  be the number

of links going out of page  $p_j$ .<sup>3</sup> Then, the PageRank of page  $p_i$  is given by

$$PR(p_i) = d + (1 - d) [PR(p_1)/c_1 + \dots + PR(p_m)/c_m]$$

Here, the constant  $d$  is called a *damping factor* whose intuition is given below. Ignoring the damping factor for now, we can see that  $PR(p_i)$  is roughly the sum of  $PR(p_j)$ ’s that point to  $p_i$ . Under this formulation, note that we construct one equation per Web page  $p_i$  with the equal number of unknown  $PR(p_i)$  values. Thus, the equations can be solved for the  $PR(p_i)$  values. This computation is typically done through iterative methods, starting with all  $PR(p_i)$  values equal to 1. At each step, the new  $PR(p_i)$  values are computed from the old  $PR(p_i)$  values from the previous step using the equation above, until the values converge.<sup>4</sup>

One intuitive model for PageRank is that we can think of a user “surfing” the Web, starting from any page, and randomly selecting from that page a link to follow.<sup>5</sup> When the user is on a page, there is some probability,  $d$ , that the next visited page will be completely random. This damping factor  $d$  makes sense because users will only continue clicking on links for a finite amount of time before they get distracted and start exploring something completely unrelated. With the remaining probability  $1 - d$ , the user will click on one of the  $c_j$  links on page  $p_j$  at random. The  $PR(p_i)$  values we computed above give us the probability that our random surfer is at  $p_i$  at any given time.

Given the definition, we can interpret the PageRank of a page as its popularity on the Web. High PageRank implies that 1) many pages on the Web are “interested” in the page and that 2) more users are likely to visit the page compared to low PageRank pages. PageRank has proven to be a very effective ranking metric for Web pages. Google was the first search engine that used PageRank as the primary ranking metric [5], but after the huge success of Google, it has been incorporated into most major search engines [21].

## 3. POPULARITY EVOLUTION: EXPERIMENTAL STUDY

We now continue our main discussion on how the popularity of Web pages evolve over time and how search engines impact the evolution. Our main goal of this section is to see whether the “rich-get-richer” phenomenon is actually happening in the current Web by conducting experiments.

In our experiments, we use both 1) the total number of incoming links to a page and 2) PageRank as the measure of popularity. To obtain these numbers, we need to know the link structure of the Web and its change over time. For this purpose, we capture two snapshots of the Web at different times, compute the PageRank (and the number of incoming links) for each page, and measure the difference between the snapshots. From this comparison, we can tell how much more popular a page has become between the snapshots. We explain our experimental setup in more detail in the next section.

<sup>3</sup>If a page has no outgoing link, we assume that it has outgoing links to every single Web page.

<sup>4</sup>This calculation corresponds to computing the principal eigenvector of the link matrix. For more details on PageRank, see reference [19].

<sup>5</sup>When the user reaches a page with no outlinks, he jumps to a random page.

### 3.1 Experimental Setup

Due to our limited network and storage resources we had to restrict our experiments to a relatively small subset of the Web. For our experiments we downloaded pages of 154 Web sites twice over a period of seven months. The list of the Web sites were collected from the Open Directory (<http://dmoz.org>). Our snapshots were complete mirrors of the 154 Web Sites. We downloaded pages from each site until we could not reach any more pages from that site or we downloaded the maximum of 200,000 pages. Only 4 Web sites had over 200,000 pages. The number of pages that were downloaded in each snapshot varied from 4.6 million to 5 million.

We analyzed the contents of these downloaded pages to obtain the outgoing links from each page. Using these outgoing links we obtain a directed graph of the Web for each snapshot. Each node in the graph corresponds to a unique Web page and an edge from the  $i^{th}$  to the  $j^{th}$  node signifies that there is an outgoing link from the  $i^{th}$  to the  $j^{th}$  Web page. While we downloaded fewer than 5 million pages in each snapshot, note that our Web graph may contain more than 5 million nodes. That is, if a page  $p_1$  (that we have downloaded) has an outgoing link to  $p_2$ , even if we have not downloaded  $p_2$ , we can still include  $p_2$  in our Web graph together with the link from  $p_1$  to  $p_2$ .<sup>6</sup> We decide to elect this option. Thus, our Web graph for the 1<sup>st</sup> snapshot contains 13 million nodes and the Web graph for the 2<sup>nd</sup> snapshot contains 15 million nodes. We will refer to our first snapshot as  $S_1$  and the second snapshot as  $S_2$ .

For each snapshot described above, we compute the **PageRank** and the total number of **incoming links** for each page. In computing PageRank, we use 0.3 as the *damping factor* (Section 2) and use 1 as the initial PageRank value of each page. Since we are interested in how the popularity of a page changes over time, we then identify the set of pages that are common in both snapshots and compare their PageRank (or the total number of incoming links) between the snapshots. There are around 7.8 million common nodes in both snapshots. The results presented below are based on these 7.8 million pages.

### 3.2 Popularity evolution

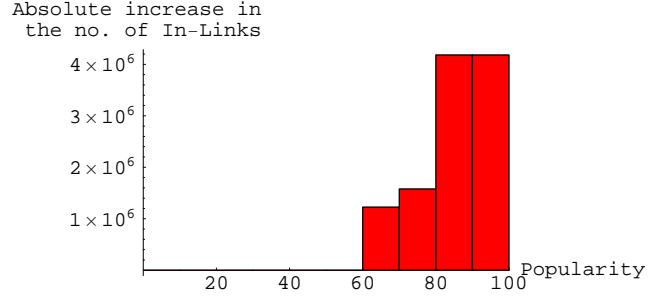
We first report our results when we use the **number of incoming links** as the **measure of popularity**. Since we are interested in knowing whether popular pages get even more popular, we divide our 7.8 million pages into ten groups based on their popularity in the first snapshot (0.78 million pages in each group). For example, we put the bottom 10% pages with the least incoming links into the first group, and the the next 10% into the second group, etc. We then examine how the popularity of each group changes between the two snapshots. If popular pages get more popular, the pages in the tenth group (top 10%) will acquire the most links.

More formally, we define the *total number of incoming links to group  $G_i$  in snapshot  $S_j$* ,  $IL(G_i, S_j)$ , as

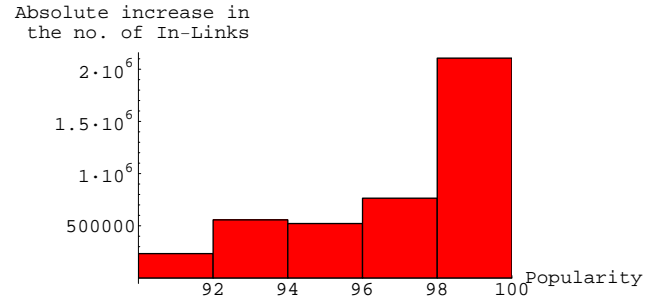
$$IL(G_i, S_j) = \sum_{p \in G_i} IL(p, S_j)$$

where  $IL(p, S_j)$  is the number of incoming links to the page

<sup>6</sup>Of course any links from  $p_2$  to other pages will be ignored in our graph because we have not downloaded  $p_2$ .



**Figure 1:** The graph shows the popularity on the X axis and the absolute change in the values of the incoming links on the Y axis.



**Figure 2:** The graph shows the popularity on the X axis and the absolute change in the values of the incoming links on the Y axis for the top 10% most popular pages.

$p$  in  $S_j$ . Then  $IL(G_i, S_2) - IL(G_i, S_1)$  represents the increase of the popularity of the group  $G_i$ .

In Figure 1, we show the popularity increase of each group. The horizontal axis represents the ten groups, where 90%–100% corresponds to the most popular group. The vertical axis shows  $IL(G, S_2) - IL(G, S_1)$  of each group. From the graph, we can clearly see that it is **only the popular pages** that become **more popular** over time. While the bottom 60% group obtained virtually no new links, the top 20% pages acquired 8 million new links (the sum of two right-most bars), which is more than 70% of all new links. In Figure 2, we show a more detailed view of the top 10% group. We further divide the top group into 5 subgroups and plot their popularity increase. Here again, we can see that the most popular pages (98%–100%) obtain significantly more new links than others.

In Figure 3, we show the *relative increase* in the popularity of each group. That is, we divide the popularity increase by current popularity (i.e.,  $[IL(G, S_2) - IL(G, S_1)] / IL(G, S_1)$ ) and plot this number. From this graph, we can see that while the pages in the 60%–80% group show a high increase rate compared to the 80%–100% group (mainly because the 60%–80% group has significantly fewer links in the first snapshot

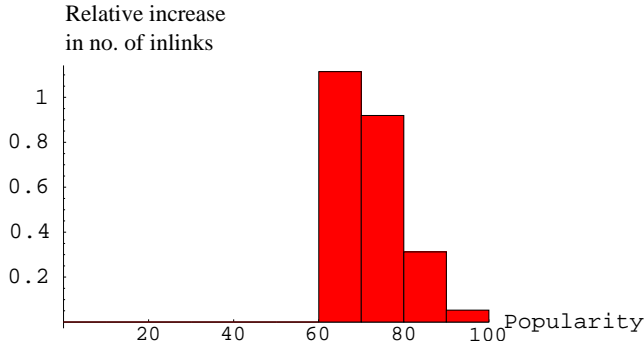


Figure 3: The graph shows the popularity on the X axis and the relative increase in the number of incoming links on the Y axis.

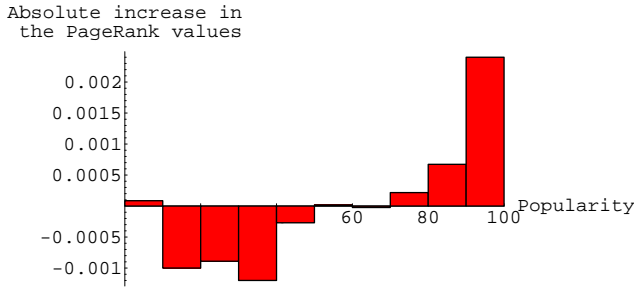


Figure 4: The graph shows the popularity on the X axis and the absolute change in the values of the PageRank on the Y axis.

than the 80%–100% group), we can still see that the bottom 60% of the pages do not show any increase in popularity. The unpopular pages are simply being ignored by users.

We obtain similar results when we use **PageRank** as the **popularity metric**. We again group the pages based on their PageRanks in the first snapshot and define the total PageRank of the group  $G_i$  in snapshot  $S_j$  as

$$PR(G_i, S_j) = \sum_{p \in G_i} PR(p, S_j).$$

Figure 4 shows the popularity increase measured in PageRank,  $PR(G, S_2) - PR(G, S_1)$ . Again, the horizontal axis represents groups and vertical axis is popularity increase. From this graph, we can see that while the pages in the 70%–100% group increase their popularity, the pages in the 20%–50% group actually decrease their popularity. That is, unpopular pages are getting even less popular! This result is mainly because PageRank is a normalized metric. PageRank measures the probability that a random Web surfer arrives at a page, so if some pages become more popular and obtain higher PageRank, then other pages should have lower PageRanks so that the overall probability is the same. In contrast, for the  $IL(p)$  metric, popularity does not decrease as long as the page does not lose its incoming links.

In Figure 5, we show a more detailed view of PageRank increase for the top 20% group. Again, we see that

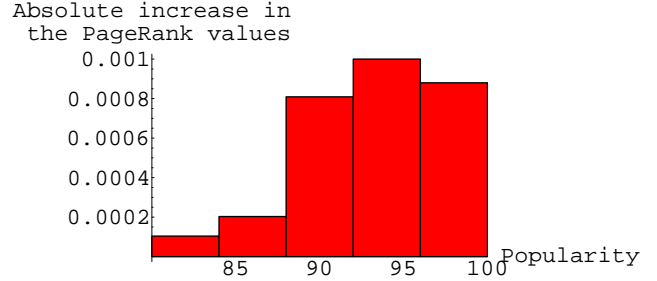


Figure 5: The graph shows the popularity on the X axis and the absolute change in the values of pagerank on the Y axis for the top 20% most popular pages.

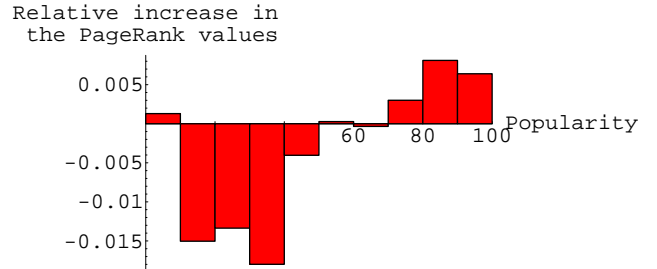


Figure 6: The graph shows the popularity on the X axis and the relative change in the values of the PageRank on the Y axis.

more popular pages tend to increase their popularity more. Finally, Figure 6 shows the relative increase of PageRank,  $[PR(G, S_2) - PR(G, S_1)] / PR(G, S_1)$ , which shows similar trend that we have observed so far.

## 4. POPULARITY EVOLUTION WITHOUT SEARCH ENGINES

In the previous section, we presented our experimental result showing that popular pages are indeed getting more popular. How much of this trend is due to the search-engine bias? That is, if search engines do not rank pages based on the current popularity, will popular pages still get more popular?

Answering this question is not easy in practice, because we cannot prevent users from using search engines in order to observe the popularity evolution when search engines do not exist. Thus, in the rest of this paper, we try to examine the impact of search engines theoretically by analyzing two Web-surfing models: the *random-surfer model* and the *search-dominant model*. The random-surfer model captures the case when the users are not influenced by search engines. In this model, Web users discover new pages simply by surf-



ing the Web, just following links. They never use a search engine to discover pages. In contrast, the search-dominant model captures the case when users' browsing patterns are completely influenced by search engines. Whenever a user wants to explore the Web, she goes to a search engine, issues queries, and clicks on the results. She never visits a page if it is not returned by a search engine.

By analyzing the popularity evolution under these two models and comparing the results, we will be able to tell how much bias search engines introduce. We first discuss the random-surfer model in this section. (The search-dominant model is discussed in Section 5.) In Section 4.1, we define the random-surfer model formally. In Section 4.2, we analyze how the popularity of a page evolves over time under the model. Then in Section 4.3, we check the validity of our random-surfer model by comparing the actual popularity evolution of Google with the result of our random-surfer model. We note that the material in Sections 4.1 through 4.3 was presented in our earlier paper [7].

#### 4.1 Random-surfer model

For our random-surfer model, we define two notions of popularity. Our first notion of popularity, (*simple*) *popularity*, measures how many Web users like a particular page.

**Definition 1 (Popularity)** We define the *popularity* of page  $p$  at time  $t$ ,  $\mathcal{P}(p, t)$ , as the fraction of Web users who like the page.  $\square$

Under this definition, if 100,000 users (out of, say, one million) currently like page  $p$ , its popularity is 0.1.

Our second notion of popularity, *visit popularity*, measures how many users visit a page at a particular time.

**Definition 2 (Visit popularity)** We define the *visit popularity* of a page  $p$  at time  $t$ ,  $\mathcal{V}(p, t)$ , as the number of "visits" or "page views" a page gets within a unit time interval at time  $t$ .  $\square$

Using these two definitions, we now introduce the two core assumptions of our random-surfer model. The first assumption of our random-surfer model is that the number of visitors to a page is proportional to its current PageRank (which can be interpreted as its popularity).

**Proposition 1** *The number of visits to page  $p$  within a unit time interval at time  $t$  is **proportional** to how many people like the page. That is,*

$$\mathcal{V}(p, t) = r_1 \mathcal{P}(p, t)$$

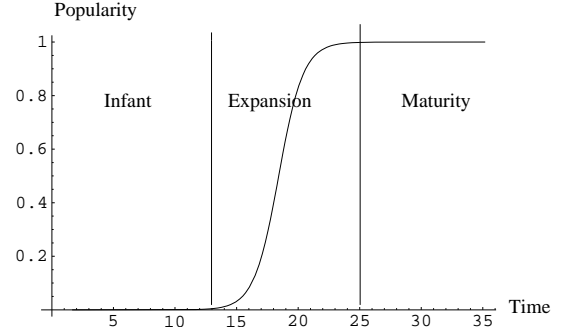
where  $r_1$  is a normalization constant.  $\square$

Intuitively, this assumption makes sense, because if a page is more popular the page is more likely to be visited. More formally, we know that the current PageRank of a page represents the probability that a person arrives at the page if the person follows links on the Web randomly (Section 2). Therefore, assuming that we use PageRank as the measure of the popularity of page  $p$ ,  $\mathcal{P}(p, t)$ , the number of visitors to the page is proportional to  $\mathcal{P}(p, t)$  under the random-surfer model.

Our second assumption is that a visit to a page can be done by any Web user with equal probability. That is, if there exist  $n$  Web users and if the page  $p$  was just visited by a user, the visit may have been done by any Web user with  $1/n$  probability.

Symbol	Meaning
$\mathcal{P}(p, t)$	(Simple) popularity of $p$ at $t$
$\mathcal{V}(p, t)$	Visit popularity of $p$ at $t$
$R(p, t)$	The ranking $p$ at $t$ in a search result
$Q(p)$	Quality of $p$
$n$	Total number of Web users
$r_1, r_2$	Normalization constants for the visit popularity and the simple popularity

**Table 1: The symbols that are used throughout this paper and their meanings**



**Figure 7: Time evolution of page popularity**

**Proposition 2** *Any visit to a page can be done by any Web user with equal probability.*  $\square$

Given these two hypotheses, we can derive how the popularity of a page evolves over time. In the next section, we present the result from this analysis. For the reader's convenience, we summarize our notation in Table 1. As we continue our discussion, we will explain some of the symbols that have not been introduced yet.

#### 4.2 Popularity evolution

Intuitively, if we know the current popularity of the page  $p$ , we can estimate how many new users will visit  $p$  based on Propositions 1 and 2. Then, if we know what fraction of these new users end up liking  $p$ , we can estimate how much its popularity increases. To capture this fraction, we define the *quality of page  $p$* ,  $Q(p)$ , as the probability that an average user will like the page  $p$  when she visits  $p$ . For example, when all users like  $p$  when they visit, the quality  $Q(p)$  is close to one.

In [7], we analyzed the popularity evolution for the random-surfer model and obtained the following result:

**Theorem 1** *The popularity of page  $p$  evolves over time through the following **formula**.*

$$\mathcal{P}(p, t) = \frac{Q(p)}{1 + \left[ \frac{Q(p)}{\mathcal{P}(p, 0)} - 1 \right] e^{-\left[ \frac{r_1}{n} Q(p) \right] t}}$$

Here,  $n$  is the total number of Web users.  $\mathcal{P}(p, 0)$  is the initial popularity of  $p$  at time zero when the page was first created.  $\square$

In Figure 7, we show an example of the time evolution of page popularity. The horizontal axis corresponds to the time. The vertical axis corresponds to the popularity  $\mathcal{P}(p, t)$  at the given time. We assumed  $Q(p) = 1$ ,  $r_1/n = 1$  and

$\mathcal{P}(p, 0) = 10^{-8}$  for this graph. These parameters represent the case when the quality of the page is very high (almost all users who look at the page like it) and the initial popularity of the page is low (only 1 out of 100 million users like the page in the beginning).

From the graph, we can see that a page roughly goes through three stages after its birth: the infant stage, the expansion stage, and the maturity stage. In the first infant stage (between  $t = 0$  and  $t = 13$ ) the page is barely noticed by Web users and has practically zero popularity. At some point ( $t = 13$ ), however, the page enters the second expansion stage ( $t = 13$  and 25), where the popularity of the page suddenly increases. Clearly, the length of the infant stage depends on the initial popularity. The higher the initial popularity is, the shorter the infant stage is. In the third maturity stage, the popularity of the page stabilizes at a certain value. The maturity stage occurs when most of the Web users have already visited the page and are aware of it.

In the next section, we compare the popularity evolution from our random-surfer model against the actual popularity evolution of a Web site to see how well they fit.

### 4.3 Case study: Google’s popularity evolution

In examining the actual popularity evolution of a Web site, there are two potential methods. First method is to examine large snapshots of the Web collected over a long period of time and investigate the link-structure changes. However, our dataset is too short for this purpose, so we cannot use this method. The second method is to use the “site-popularity” data reported by Web-rating companies. For example, Nielsen-NetRatings [17] tracks how many Web users visit some of the well-known Web sites in each week and publishes their finding every week. Since this data is available as early as 1996, we decided to use this method.

For our comparison, we use Google’s popularity evolution, because it is one of the few companies that Nielsen-NetRatings (and other Web-rating companies) started to track from the beginning of the company. Other popular Web sites, such as Yahoo and AOL, went online much before Nielsen-NetRatings began tracking them, so we do not know their complete popularity evolution. In addition, we believe that Google is the Web site which is *least* affected by popularity-based ranking mechanisms. Google is the **first** search engine that **used PageRank** as their main ranking function and PageRank had not been implemented by other search engines for a few years. Therefore, Google’s popularity evolution had **not** been **affected** by popularity-based rankings from other search engines — at least initially until other search engines started to implement variations of PageRank.

Roughly speaking, Nielsen-NetRatings tracks what fraction of Web users visit each Web site every week, by installing their monitoring program on a number of computers and tracking them constantly. Each computer is used exclusively by a single person, so one machine corresponds to one Web user. From this tracking, Nielsen reports what they call the *audience reach*, which is the fraction of their Web users who visit a particular site at least once in each week. For example, the audience reach 0.3 means that 30% of the users visited the site at least once in the week.

We downloaded Google’s audience-reach data from Nielsen-NetRatings and plotted graph in Figure 8. The solid line

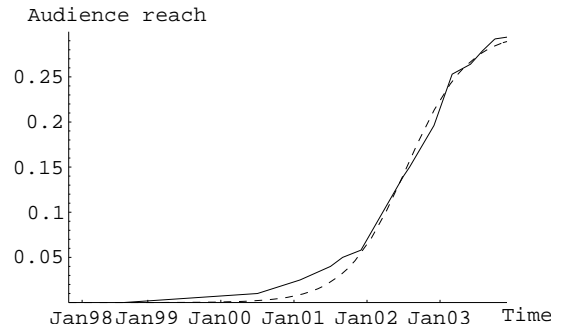


Figure 8: Google’s popularity evolution

in the graph is the Nielsen’s experimental data. The graph starts at January 1998 when Google’s first prototype system went online. From the graph, we observe that Google had been relatively unknown until mid-2000, but from that point on its popularity exploded. This explosion lasted for about three years until early 2003 when the popularity started to level off at 0.3.

For comparison, we show the prediction of our random-surfer model as a dashed line in the graph, with the following parameters:  $Q(p) = 0.3$ ,  $\mathcal{P}(p, 0) = 5 \times 10^{-6}$ , and  $\frac{c_1}{n} = 8$ . These parameters mean that an average Web user likes Google with a 30% probability and that one out of 200,000 Web users initially liked Google when it went online. The initial popularity is rather high given that there were more than 200,000 users in 1998. Perhaps this high initial popularity may be due to the “Stanford aura effect.” Since Google was linked from many Stanford pages and because a large number of Web users were visiting Stanford sites regularly after a number of successful Internet companies started from Stanford, Google may have attracted relatively large initial traffic compared to an average Web site.

In general, we can see that Google’s popularity evolution follows an S-curve as predicted by our model. Given this result, we believe that our random-surfer model captures the popularity evolution of Web pages reasonably well.

## 5. IMPACT OF SEARCH ENGINES ON POPULARITY EVOLUTION

In the previous section, we studied the popularity evolution of a page when users discover pages purely based on random surfing. In this section, we analyze how the popularity evolution changes when the users discover pages solely based on search results (the search-dominant model). From this analysis, we can tell how long it takes for a page to become popular when users’ browsing pattern is dominated by a search engine, and thus we can indirectly measure the potential bias introduced by search engines. As we will see later, the result from our analysis is quite alarming. Our result predicts that it takes 66 times longer under the search-dominant model than under the random-surfer model in order for a page to become popular! This result strongly indicates that we need to devise a new mechanism to “promote” new pages, so that new pages have higher chance to be “discovered” by people and get the attention that they may deserve. We further discuss this issue at the end of this section.

## 5.1 Search-dominant model

For our search-dominant model, we assume that the users use only one search engine. In addition, we assume that the search engine always returns the *same* set of pages in the *same* order, ranked purely by their popularity. This assumption may be unrealistic for general search scenarios because search engines return *different* results depending on the query. However, we may consider that our model investigates the set of pages returned for a particular query, say, “XML.” For all “XML” queries, the search engine returns the same set of pages (related to XML) and ranks the pages roughly by their PageRank. Therefore, if we focus our attention only to this set of pages, their relative popularity evolution will be similar to what our search-dominant model predicts.

In formalizing our search-dominant model, we first note that the main assumption for the random-surfer model is Proposition 1: the visit popularity of a page is proportional to its current popularity. This assumption makes sense when users surf the Web randomly (Section 2), but it may not be valid when users visit pages purely based on search results. Then what will be a good model to estimate the visit popularity?

We can derive the relationship between  $\mathcal{V}(p, t)$  and  $\mathcal{P}(p, t)$  by investigating the following two distributions:

1. If a page is returned as the  $i^{th}$  entry in the search result, how likely is the user to click on the page? For example, what fraction of users will visit the second entry in the search result?
2. Given the PageRank of a page, what will be its ranking in the search result?

In Section 5.2, we present the empirical data that provides the answers to the above questions. Based on this empirical data, we assume the following relationship between  $\mathcal{V}(p, t)$  and  $\mathcal{P}(p, t)$  for the search-dominant model.

**Proposition 3** *Under the search-dominant model, the number of visits to page  $p$  at time  $t$  satisfies the following equation:*

$$\mathcal{V}(p, t) = r_2 \mathcal{P}(p, t)^{\frac{9}{4}}$$

where  $r_2$  is a normalization constant.  $\square$

Under this hypothesis, note that users visit popular pages significantly more often than unpopular pages compared to the random-surfer model. We illustrate this point through an example.

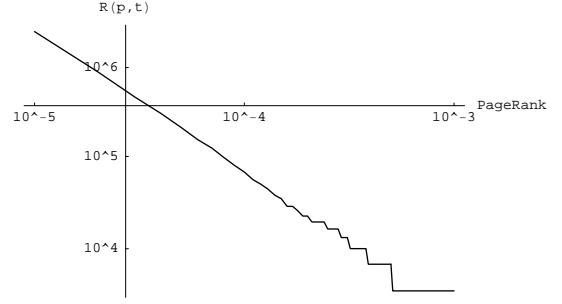
**Example 1** Consider two pages,  $p_1$  and  $p_2$ , with popularity values at 0.9 and 0.1, respectively. Under the random-surfer model, the visit popularity is proportional to the popularity, so

$$\frac{\mathcal{V}(p_1, t)}{\mathcal{V}(p_2, t)} = \frac{\mathcal{P}(p_1, t)}{\mathcal{P}(p_2, t)} = \frac{0.9}{0.1} = 9.$$

That is,  $p_1$  is visited 9 times more often than  $p_2$ . Under the search-dominant model, however, the visit popularity is proportional to  $\mathcal{P}(p, t)^{\frac{9}{4}}$ , so

$$\frac{\mathcal{V}(p_1, t)}{\mathcal{V}(p_2, t)} = \left[ \frac{\mathcal{P}(p_1, t)}{\mathcal{P}(p_2, t)} \right]^{\frac{9}{4}} = \left[ \frac{0.9}{0.1} \right]^{\frac{9}{4}} = 140.$$

That is,  $p_1$  is visited 140 times more often than  $p_2$ .  $\square$



**Figure 9: Probabilistic cumulative distribution of PageRank values**

This result is reasonable. Since the search engine mainly “promotes” popular pages by returning them at the top, they are visited more often than under the random-surfer model.

In the next section, we present empirical evidences that lead to Proposition 3. A reader who is not interested in the **derivation** of Proposition 3 may skip the next section.

## 5.2 Visit popularity under the search-dominant model

In the search-dominant model, users surf the Web starting from the search result page and clicking on the returned URLs. In addition, they are more likely to click on the top-result pages than the bottom ones. Therefore, the probability to visit page  $p$ ,  $\mathcal{V}(p, t)$  depends on the rank of  $p$  in the search result. We use  $R(p, t)$  to represent the **rank** of  $p$  at time  $t$ . For example, if  $p$  is the  $2^{nd}$ -ranked page,  $R(p, t)$  is 2. Then how likely are the users to click on the  $i^{th}$  result? That is, what is the relationship between  $R(p, t)$  and  $\mathcal{V}(p, t)$ ?

Wolf et al. [23] proposed a mathematical model for the users’ click probability of a page and the rank of the page. Lempel and Moran [16] provide empirical measurements of the click probability and the rank of a page from the AltaVista query log. According to their empirical measurements, it appears that the click probability closely follows the distribution below:

$$\mathcal{V}(p, t) = c_1 R(p, t)^{-\frac{3}{2}} \quad (1)$$

Here,  $c_1$  is a normalization constant. Based on this empirical data, we will assume Equation 1 as the relationship between  $\mathcal{V}(p, t)$  and  $R(p, t)$ .

We next examine the relationship between  $R(p, t)$  and  $\mathcal{P}(p, t)$ . To derive this relationship, we note that  $R(p, t)$  is the rank of  $p$  when pages are ordered by their PageRank values. That is,  $R(p, t)$  is equivalent to the number of pages whose PageRank values are above  $\mathcal{P}(p, t)$ .<sup>7</sup> Therefore, if we know the overall distribution of PageRank values, we can obtain the rank of  $p$ ,  $R(p, t)$ , from its PageRank value,  $\mathcal{P}(p, t)$ .

In Figure 9, we show the PageRank distribution obtained from a snapshot of the Web. The distribution was obtained from a Web snapshot captured by Stanford WebBase project [13]. The WebBase project periodically downloads hundreds of millions of pages on the Web, stores the pages in their local repository and provides them to researchers in

<sup>7</sup>More precisely, this number plus one

other institutions. The graph in Figure 9 was obtained from a snapshot containing roughly 100 million pages. The horizontal axis in the graph corresponds to the PageRank value and the vertical axis shows the ranking of the pages at the given PageRank value. Both axes in the graph are in the logarithmic scale. Since the graph is a straight line in the logarithmic scale with the slope  $-\frac{3}{2}$ , we see that the ranking and the PageRank of a page satisfy following equation:

$$R(p, t) = c_2 \mathcal{P}(p, t)^{-\frac{3}{2}} \quad (2)$$

where  $c_2$  is a normalization constant.

While we obtained Figure 9 from a snapshot of the WebBase repository, this PageRank distribution seems to be universally true on the Web. We observed almost identical distributions for other snapshots in the WebBase repository and a roughly equivalent distribution from the two snapshots described in Section 3. Given these results, we assume the PageRank distribution follows Equation 2.

Given Equations 1 and 2, we can derive the following relationship between  $\mathcal{V}(p, t)$  and  $\mathcal{P}(p, t)$ :

$$\begin{aligned} \mathcal{V}(p, t) &= c_1 R(p, t)^{-\frac{3}{2}} \\ &= c_1 \left( c_2 \mathcal{P}(p, t)^{-\frac{3}{2}} \right)^{-\frac{3}{2}} \\ &= r_2 \mathcal{P}(p, t)^{\frac{9}{4}} \end{aligned}$$

### 5.3 Popularity evolution

In the previous section, we explained the main hypothesis of the search-dominant model, Proposition 3, that shows how visit popularity is related to the simple popularity. In addition to this hypothesis, if we assume Proposition 2 (the visits to a page are done by random users), we can analyze the popularity evolution for the search-dominant model. The following theorem is the result of this analysis.

**Theorem 2** *Under the search-dominant model, the popularity of page  $p$ ,  $\mathcal{P}(p, t)$ , evolves through the following equation:*

$$\sum_{i=1}^{\infty} \frac{[\mathcal{P}(p, t)]^{(i-\frac{9}{4})} - [\mathcal{P}(p, 0)]^{(i-\frac{9}{4})}}{(i-\frac{9}{4}) Q(p)^i} = \frac{r_2}{n} t$$

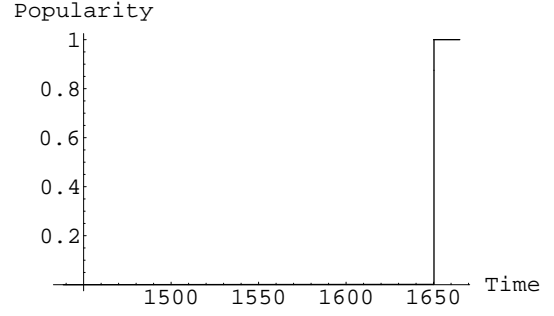
Here,  $n$  is the total number of Web users and  $\mathcal{P}(p, 0)$  is the initial popularity of  $p$  when the page was first created.  $\square$

We defer the proof of the above theorem to Section 9, and first study its implication.

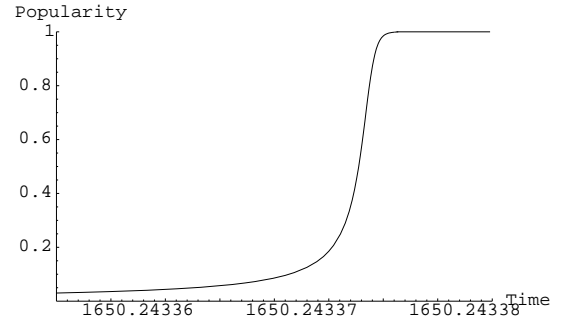
In Figure 10, we show the popularity evolution of a page under the search-dominant model for the **same parameters** as in Figure 7.<sup>8</sup> Figure 11 shows the same graph, but only around  $t = 1650$ , when the popularity suddenly increases. In both graphs, the horizontal axis is time and the vertical axis is  $\mathcal{P}(p, t)$ . The figures show an alarming impact of a search engine on page popularity.

1. It takes **several orders** of **magnitude** more **time** for a page to become popular under the search-dominant model. In Figure 7 (random-surfer model), it took less

<sup>8</sup>For a fair comparison, we should set  $r_2$  such that the total number of visits to overall pages are the same under both models. That is,  $\sum_{i=1}^n \mathcal{V}(p_i, t) = \sum_{i=1}^n r_1 \mathcal{P}(p_i, t) = \sum_{i=1}^n r_2 \mathcal{P}(p_i, t)^{\frac{9}{4}}$ . We set  $r_2 = 4.836 \times 10^6 r_1$  so that this equation holds.



**Figure 10: Popularity evolution under the search-dominant model**



**Figure 11: Closer look at the popularity evolution around  $t = 1650$**

than **25** time units for the page to obtain popularity one, but in Figure 10 (search-dominant model), it took **1650** time units! It means that if a page becomes popular within one year when search engines do not exist, it takes 66 years when search engines dominate users' browsing pattern!

2. The popularity increase is much **more sudden** under the search-dominant model than under the random-surfer model. In Figure 7, the page spent around 12 time units (from  $t = 13$  until  $t = 25$ ) in the expansion stage, but the expansion stage virtually does not exist in Figure 10; the popularity increases from zero to one almost immediately at  $t = 1650$ .

This result is **because** most of the **user traffic** is **directed** to **popular** pages under the search-dominant model. Therefore, unpopular pages get significantly less traffic than under the random-surfer model, so it takes much longer time for a page to build up initial momentum. However, once it obtains a reasonable ranking in the search result, it garners significantly more traffic than under the random-surfer model, so its popularity increases very quickly as long as it is of high quality.

Given this result, we believe that search engines can play a very significant role in the **"survival"** of a page. As Figure 10 shows, once a page starts to get noticed by Web users, its popularity can jump almost immediately (as long as the page is of high quality). Thus, if search engines can **identify** high **quality** pages early on and promote them for a relatively short period, the pages can achieve its eventual popularity significantly earlier than under the random-surfer model.



## 6. RELATED WORK

There have been a number of studies investigating the evolution of the Web pages [4, 8, 10, 11, 22]. Most of these studies are experimental and mainly focus on the changes in the content of the pages. For example, Fetterly et al. [11] downloaded close to 100 million pages on a weekly basis for 11 weeks and examined how often the pages changed, how significant changes occurred and what were the major factors influencing the degree of change of each page.

Link-popularity metrics were first proposed by Kleinberg [15] and Page et al. [19]. Google is the first company that adopted a link-popularity metric as the primary ranking metric [5]. Major search engines have adopted variations of PageRank in the last few years after Google became hugely successful [21]. In this paper, we study the impact of link-popularity metrics on the popularity evolution of Web pages. In our related paper [7], we propose a new ranking metric that discovers “high-quality” pages early on, so that we can minimize the popularity bias introduced by search engines.

There exists a large body of work that investigates the properties of the Web link structure [1, 2, 6, 20]. For example, [6] shows that the global link structure of the Web is similar to a “bow-tie.” [1, 6] show that the number of incoming or outgoing links follow a power-law distribution. [2, 20] propose potential models on the Web link structure. Note that these studies investigate the distribution of links within a *snapshot* of the Web, while we study the popularity evolution over time.

There exist a number of studies that measure the user traffic to individual Web sites and/or pages [3, 14, 9]. In most of the study, the traffic also seems to follow the power-law distribution. [14] proposes a theoretical model to explain the overall traffic distribution. Again, our work focuses on the time evolution of popularity rather than the overall traffic distribution at a particular point in time.

## 7. CONCLUSION

In this paper, we studied how the popularity of a Web page evolves over time and how search engines affect the popularity evolution. Through an **experimental** study conducted over 7 months, we first showed that popular pages are indeed getting more popular while unpopular pages are getting relatively less popular. We then analyzed two reasonable Web **models** and tried to estimate the potential impact of search engines on the popularity evolution of Web pages. The result from our analysis is immensely **worrisome**. It shows that when search engines rank pages based on their popularity, it takes several **orders** of **magnitude** more **time** for a new page to become popular even if the page is of high quality. Given that PageRank and its variations are being used by major search engines, our result strongly indicates that many high-quality pages are ignored by Web users, simply because no one has discovered them yet. We believe that our study demonstrates an urgent need to develop a new ranking mechanism (such as the one proposed in [7]) that can potentially identify high-quality pages early on and promote them, so that we can alleviate this problem.

## 8. REFERENCES

- [1] R. Albert, A.-L. Barabasi, and H. Jeong. Diameter of the World Wide Web. *Nature*, 401(6749):130–131, September 1999.
- [2] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.
- [3] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in web client access patterns: Characteristics and caching implications. *World Wide Web*, 2(1-2):15–28, 1999.
- [4] B. E. Brewington and G. Cybenko. How dynamic is the web. In *Proceedings of the International World-Wide Web Conference*, May 2000.
- [5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the International World-Wide Web Conference*, April 1998.
- [6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. In *Proceedings of the International World-Wide Web Conference*, May 2000.
- [7] J. Cho and R. E. Adams. Page quality: In search of an unbiased web ranking. Technical report, UCLA Computer Science Department, November 2003.
- [8] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *Proceedings of the Twenty-sixth International Conference on Very Large Databases (VLDB)*, Cairo, Egypt, September 2000.
- [9] B. D. Davison. Web traffic logs: An imperfect resource for evaluation. In *Ninth Annual Conference of the Internet Society (INET’99)*, pages 22–25, June 1999.
- [10] F. Douglass, A. Feldmann, and B. Krishnamurthy. Rate of change and other metrics: a live study of the world wide web. In *Proceedings of the Second USENIX Symposium on Internetworking Technologies and Systems*, October 1999.
- [11] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the International World-Wide Web Conference*, pages 669–678, 2003.
- [12] Google information for webmasters. Available at <http://www.google.com/webmasters/>.
- [13] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. Webbase: A repository of web pages. In *Proceedings of the International World-Wide Web Conference*, pages 277–293, May 2000.
- [14] B. A. Huberman, P. Pirollo, J. Pitkow, and R. M. Lukose. Strong regularities in world wide web surfing. *Science*, 280(5360):95–97, April 1998.
- [15] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.
- [16] R. Lempel and S. Moran. Predictive caching and prefetching of query results in search engines. In *Proceedings of the International World-Wide Web Conference*, May 2003.
- [17] Nielsen netratings search engine ratings. Brief summary is available at <http://searchenginewatch.com/reports/article.php/2156451>, February 2003.
- [18] S. Olsen. Does search engine’s power threaten web’s independence? Available at <http://news.com.com/2009-1023-963618.html>,

October 2002.

- [19] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University Database Group, 1998. Available at <http://dbpubs.stanford.edu:8090/pub/1999-66>.
- [20] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211, 2002.
- [21] J. A. Tomlin. A new paradigm for ranking pages on the world wide web. In *Proceedings of the International World-Wide Web Conference*, May 2003.
- [22] C. E. Wills and M. Mikhailov. Towards a better understanding of web resources and server responses for improved caching. In *Proceedings of the International World-Wide Web Conference*, May 1999.
- [23] J. L. Wolf, M. S. Squillante, P. S. Yu, J. Sethuraman, and L. Ozsen. Optimal crawling strategies for web search engines. In *Proceedings of the International World-Wide Web Conference*, May 2002.

## 9. DERIVATION OF POPULARITY EVOLUTION

In this section, we provide the detailed analysis of the popularity evolution discussed in Section 5.3. The analysis of the random-surfer model was provided in our earlier paper [7].

To help our analysis, we first define the *user awareness* of page  $p$  at time  $t$ ,  $\mathcal{A}(p, t)$ , as the fraction of the Web users who is aware of the page at the time. For example, if 100,000 users (say, out of one million) have visited the page  $p_1$  so far and are aware of the page, its user awareness,  $\mathcal{A}(p_1, t)$ , is 0.1. Given the definition, we can see that we can measure the popularity of a page from its awareness and quality.

$$\mathcal{P}(p, t) = \mathcal{A}(p, t) \cdot Q(p) \quad (3)$$

The above equation holds because the Web users who currently like the page  $p$  are the people who are aware of the page *and* like it.

Assuming that there are  $n$  Web users in total, we now derive how the current awareness of a page is related to its past popularity.

**Lemma 1** *The user awareness of  $p$  at  $t$ ,  $\mathcal{A}(p, t)$ , can be computed from its past popularity through the following formula:*

$$\mathcal{A}(p, t) = 1 - e^{-\frac{r_2}{n} \int_0^t [\mathcal{P}(p, t)]^k dt} \quad (4)$$

Here,  $k$  is the constant  $\frac{9}{4}$  which we assumed in Proposition 3.  $\square$

**Proof**  $\mathcal{V}(p, t)$  is the rate at which Web users visit the page  $p$  at  $t$ . Thus by time  $t$ , page  $p$  is visited  $\int_0^t \mathcal{V}(p, t) dt = r_2 \int_0^t [\mathcal{P}(p, t)]^k dt$  times.

Without loss of generality, we compute the probability that user  $u_1$  is not aware of the page  $p$  when the page has been visited  $m$  times. The probability that the  $i$ th visitor to  $p$  was not  $u_1$  is  $(1 - \frac{1}{n})$ . Therefore, when  $p$  has been visited  $k$  times, the probability that  $u_1$  would have never visited  $p$  is  $(1 - \frac{1}{n})^m$ . By time  $t$ , the page is visited  $\int_0^t \mathcal{V}(p, t) dt$  times.

Then the probability that the user is not aware of  $p$  at time  $t$ ,  $1 - \mathcal{A}(p, t)$ , is

$$\begin{aligned} 1 - \mathcal{A}(p, t) &= \left(1 - \frac{1}{n}\right)^{\int_0^t \mathcal{V}(p, t) dt} \\ &= \left(1 - \frac{1}{n}\right)^{r_2 \int_0^t [\mathcal{P}(p, t)]^k dt} \\ &= \left[\left(1 - \frac{1}{n}\right)^{-n}\right]^{-\frac{r_2}{n} \int_0^t [\mathcal{P}(p, t)]^k dt} \end{aligned}$$

When the number of web users is large, we can approximate the above expression by limiting  $n$  to infinity:  $n \rightarrow \infty$ ,  $(1 - \frac{1}{n})^{-n} \rightarrow e$ . Thus,

$$1 - \mathcal{A}(p, t) = e^{-\frac{r_2}{n} \int_0^t [\mathcal{P}(p, t)]^k dt} \quad \blacksquare$$

Based on Equations 3 and 4, we now derive the popularity evolution of a page.

**Proof for Theorem 2** From Equations 3 and 4,

$$\mathcal{P}(p, t) = \left[1 - e^{-\frac{r_2}{n} \int_0^t [\mathcal{P}(p, t)]^k dt}\right] Q(p) \quad (5)$$

If we differentiate both sides of the above equation,

$$\frac{d\mathcal{P}}{dt} = \left(-\frac{r_2}{n} \mathcal{P}^k\right) \left(-e^{-\frac{r_2}{n} \int_0^t \mathcal{P}^k dt}\right) Q \quad (6)$$

From Equation 5, we know that  $e^{-\frac{r_2}{n} \int_0^t \mathcal{P}^k dt} = 1 - \frac{\mathcal{P}}{Q}$ . Thus, Equation 6 becomes

$$\frac{d\mathcal{P}}{dt} = \left(\frac{r_2}{n} \mathcal{P}^k\right) \left(1 - \frac{\mathcal{P}}{Q}\right) Q.$$

After rearrangement, we get

$$\frac{1}{\left(1 - \frac{\mathcal{P}}{Q}\right) Q \mathcal{P}^k} d\mathcal{P} = \frac{r_2}{n} dt. \quad (7)$$

Since  $\frac{\mathcal{P}}{Q} < 1$ , we can use the expansion

$$\left(1 - \frac{\mathcal{P}}{Q}\right)^{-1} = 1 + \frac{\mathcal{P}}{Q} + \left(\frac{\mathcal{P}}{Q}\right)^2 + \dots = \sum_{i=0}^{\infty} \left(\frac{\mathcal{P}}{Q}\right)^i.$$

Then Equation 7 becomes

$$\left[\sum_{i=0}^{\infty} \frac{\mathcal{P}^{i-k}}{Q^{i+1}}\right] d\mathcal{P} = \frac{r_2}{n} dt.$$

If we integrate both sides of the above equation,

$$\left[\sum_{i=1}^{\infty} \frac{\mathcal{P}^{i-k}}{(i-k) Q^i}\right] + C = \frac{r_2}{n} t$$

where  $C$  is a constant determined by the boundary condition. When  $t = 0$ , the right-hand side of the above equation is zero, so the left-hand side should also be zero. Therefore,

$$\left[\sum_{i=1}^{\infty} \frac{\mathcal{P}(p, t)^{i-k} - \mathcal{P}(p, 0)^{i-k}}{(i-k) Q(p)^i}\right] = \frac{r_2}{n} t. \quad \blacksquare$$