Page Quality: In Search of an Unbiased Web Ranking

Junghoo Cho Sourashis Roy Robert E. Adams
UCLA Computer Science
{cho,roys,readams}@cs.ucla.edu

ABSTRACT

In a number of recent studies [4, 8] researchers have found that because search engines repeatedly return currently popular pages at the top of search results, popular pages tend to get even more popular, while unpopular pages get ignored by an average user. This "rich-get-richer" phenomenon is particularly problematic for new and high-quality pages because they may never get a chance to get users' attention, decreasing the overall quality of search results in the long run. In this paper, we propose a new ranking function, called *page quality* that can alleviate the problem of popularity-based ranking. We first present a formal framework to study the search engine bias by discussing what is an "ideal" way to measure the intrinsic quality of a page. We then compare how PageRank, the current ranking metric used by major search engines, differs from this ideal quality metric. This framework will help us investigate the search engine bias in more concrete terms and provide clear understanding on why PageRank is effective in many cases and exactly when it is problematic. We then propose a practical way to estimate the intrinsic page quality to avoid the inherent bias of PageRank. We derive our proposed quality estimator through a careful analysis of a reasonable web user model and we present experimental results that show the potential of our proposed estimator. We believe that our quality estimator has the potential to alleviate the rich-get-richer phenomenon and help new and high-quality pages get the attention that they deserve.

1. INTRODUCTION

Recent studies show that search engines play an increasingly important role in people's surfing of the web; when a user wants to look up information from the web, the user often goes to his favorite search engine, issues keyword queries, and clicks on 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 narrow the focus to a small number of pages worth looking at.

Given this dominant role that search engines play in our daily web access, it is now even claimed that "if your page is not indexed by Google, your page does not exist on the web [20]." While this statement may be an exaggeration, it has an alarming bit of truth. Given that an individual does not have the time to look at all web pages and identify relevant ones, a page that is not returned by search engines is unlikely to be viewed by many web users. In short, what people perceive from the web may not necessarily be what exists on the real web, but what is processed and presented by search engines.

Recently, this potential "bias" introduced by search engines on the users' perception of the web has attracted significant attention from the research community and has become an active area of research [4, 8]. For example, Cho et al. [8] studied the property of PageRank, one of the core ranking metrics used by web search engines, and presented a set of evidences that the PageRank metric induces the "rich-get-richer" phenomenon. That is, popular pages (the pages with high PageRank values) get even more popular over time because search engines repeatedly return them at the top of search results and induce more people to visit them. In contrast, a newly-created page gets completely ignored by users even if the page is of very high quality because it is ranked at the bottom of search results. In their study of Chilean web sites, Baeza-Yates et al. [4] experimentally show that new pages have significantly lower PageRank values than older ones and show that PageRank does not work well in identifying new and high-quality pages.

The main goal of this paper is to understand the limitation of PageRank and to design a new ranking metric that can alleviate the bias of PageRank. Towards this goal, we first explore what might be a good way of measuring the true "quality" of a page and how PageRank is related to this hypothetical quality metric. We then propose a practical *quality estimator* that predicts the true quality value of a page based on the evolution of the link structure of the web. As we will see, our proposed quality estimator has a strong theoretical foundation — it is derived through a careful analysis of a reasonable web user model — and can be viewed as an "improved" version of PageRank. It considers both the *current* popularity of a page and its relative popularity *increase* in measuring the quality of a page. We also present an experimental evidence that suggests the effectiveness of our quality estimator in a real-world setting. In

SIGMOD 2005 June 14-16, 2005, Baltimore, Maryland, USA.
Copyright 2005 ACM 1-59593-060-4/05/06 $5.00.

summary, we believe we make the following contributions in this paper:

- We introduce a formal definition of *page quality*, which we believe is a good way of capturing the intuitive concept of "page quality." By separating the notion of page quality from actual ranking functions, such as PageRank, we provide the formal framework to objectively judge the effectiveness of a ranking function. (Section 4)

- By comparing our definition of page quality and PageRank, we provide a formal justification on why PageRank is an effective ranking metric in many scenarios. We also show that PageRank is biased against unpopular pages, especially the ones that were created recently. (Section 4)

- We propose a direct and practical way of estimating page quality. Our proposed *quality estimator* is based on our careful analysis of a simple and reasonable web user model. (Sections 5 and 6)

- We conduct an experiment on real-world web data to measure the effectiveness of our quality estimator. While preliminary, this experiment will show the potential of our estimator in estimating the quality of a page. (Section 8)

2. RELATED WORK

[25] provides a good overview of the work done in the Information Retrieval (IR) community that studies the problem of identifying the best matching documents to a user query. This body of work analyzes the *content* of the documents to find the best matches. The boolean model [31], the vector-space model [24] and the probabilistic model [23, 9] are some of the well known models developed in this context. Some of these models (particularly the vector-space model) were adopted by most web search engines to find relevant documents to a given query. PageRank and our proposed quality estimator is applied to the set of relevant pages discovered using these models in order to rank the pages.

A number of researchers have investigated using the link structure of the web to improve search results and proposed various ranking metrics. Hub and Authority [16] and PageRank [21] are the most well known metrics that use the web link structure. PageRank and its variations are currently being used by major search engines. [1, 14, 15] describe various ways to improve PageRank computation. [2] provides a theoretical justification for the Hub and Authority metric and proposes a mechanism to combine link and text analysis for page ranking. [13] studies personalization of the PageRank metric by giving different weights to pages. [28] proposes a modification of PageRank equation to tailor it for web administrators. [30] describes how to compute the global PageRank based on local link structure within each site and the inter-site link information. [26] proposes to rank web pages by the user traffic to the pages and suggests a traffic-prediction model based on entropy maximization. In the database community, researchers also developed ways to rank database objects by modeling the object relationship as a graph and use the graph structure to rank them [11, 10, 5].

There exists a large body of work that investigates the properties of the web link structure [3, 6, 7, 22]. For example, [7] shows that the global link structure of the web is similar to a "bow tie." [3, 7] shows that the number of in-bound or out-bound links follow a power-law distribution. [6, 22] propose potential models on the web link structure.

[4, 8] provide experimental evidences that PageRank is biased against new pages. In their study of Chilean web sites, Baeza-Yates et al. [4] show that new pages have significantly lower PageRank values than others and propose to consider the last-modified date of a page in measuring the quality of a page.

The probabilistic model [9, 23] developed in the IR community is similar to our quality metric in that both definitions take a probabilistic approach. The probabilistic model, however, measures the probability that a page belongs to the relevant set given a particular user query, while our quality metric measures the general probability that a user will like a page when the user looks at the page.

3. PAGERANK AND POPULARITY

We start our discussion with a brief overview of the PageRank metric and explain how it is related to the notion of the popularity of a page. 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).

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 pages $p_1, \ldots, p_n$, which link to a page p_i. Let c_j be the total number of links going out of page p_j. If a page has no outgoing link, we assume that it has outgoing links to every single web page. Then, the PageRank of page p_i is given by

$$PR(p_i) = d + (1-d)\,[PR(p_1)/c_1 + \cdots + PR(p_n)/c_n]$$

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, 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.

A way to think intuitively about PageRank is to consider a user "surfing" the web, starting from any page, and randomly selecting from that page a link to follow. When the user reaches a page with no outlinks, he jumps to a random page. 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 web users are interested in the page and that (2) more users are likely to visit the page compared to low PageRank pages. Given the effectiveness of Google's search results and its adoption by many web search engines [26], PageRank seems to capture the importance or the quality of web pages well. According to a recent survey the majority of users are satisfied with the top-ranked results from Google and from major search engines [19].

4. QUALITY AND PAGERANK

In the previous section, we went over the definition of PageRank and explained that the PageRank of a page captures the popularity of the page on the web. We also argued that the widespread use of PageRank for web search engines indicates its effectiveness for web searches.

Before we discuss the weaknesses of PageRank and devise an improved quality metric, we first examine why PageRank is effective in ranking web pages, so that we can build our new metric on the strength of PageRank. The key feature of PageRank is that it is based on the popularity of a web page. In order for a page to be popular, may users must have examined the page *and* liked it. Given this fact, when the PageRank of a page is high — meaning that many previous users looked at the page and liked it — it is reasonable to expect that a new user seeing the page for the first time will also like it. Overall, by returning high PageRank pages first in their search results, search engines increase the *probability* that their users like their first few results.

At the same time, PageRank is significantly biased against unpopular pages, especially the ones that were recently created [8, 4]. For example, consider a new page that has just been created. We assume that the page is of very high quality and anyone who looks at the page agrees that the page should be ranked highly by search engines. Even so, because the page is new, few people are aware of it and there exist only a few (or no) links to it. This low popularity means the page will be ranked at the bottom of search results, which in turn means that few users will ever see the page. Because of the low traffic, it takes a very long time for the page to become popular.

Given the above discussion, we argue that what we really want to use as the ranking metric is *not* the current popularity of the page, but the *probability* that a web user will like the page when the user sees it for the first time. PageRank works well in many cases because it captures this probability well for well-known pages. At the same time, it is biased for new pages because PageRank does not correlate with this probability for new pages. To avoid this bias while preserving the strength of PageRank, we propose to use the following definition as the quality of a page:

Definition 1 (Page quality) We define the *quality* of a page p, $Q(p)$, as the conditional probability that an average user will like the page when user sees the page for the first time. Mathematically,

$$Q(p) = P\left(L_p | A_p\right)$$

where A_p represents the event that the user becomes newly aware of the page p by visiting the page for the first time and L_p represents the event that the user likes the page. □

Given this definition, we can hypothetically measure the quality of page p by showing p to *all* web users. For example, assuming the total number of web users is 100, if 90 web users like page p after they read it, its quality $Q(p)$ is 0.9. We discuss how we may measure page quality without explicit user feedback in the next section.

We believe that our quality definition is reasonable given that page quality can be a very subjective notion [17, 12]; one person may regard a page very highly while another person may consider the page completely useless. When individual users have different opinions on the quality of a page, it is reasonable to prefer the one that people are most likely to "vote for."

Note that PageRank of a page estimates the quality of a page well if all web pages have been given the same chance to be discovered by web users; when pages have been looked at by the same set of people, its popularity or the number of people who like the page is proportional to its quality. However, new pages have not been given the same chance as old and established pages, so the current popularity of new pages are definitely lower than their quality.

Finally, under our definition, we note that it is possible that page p_1 is considered of higher quality than p_2 simply because p_1 discusses a more popular topic. For example, if p_1 is about the movie "Star Wars" and p_2 is about the movie "Latino" (a 1985 movie produced by George Lucas), p_1 may be considered of higher quality simply because the movie "Star Wars" is more popular than "Latino." We believe this "topic bias" is not important in our context. Before search engines use our quality metric (or PageRank) to rank pages, they first use a *relevance metric* (such as the *tf.idf* metric [25]) to select the set of pages relevant to the query issued by the user. It is only within this set of pages (say, pages on the movie Latino) that the quality metric is applied. Therefore, only the relative quality within a particular relevant set of documents will actually be important in determining the results returned in response to a query. Thus, the absolute difference in the quality value among the documents on different topics does not hurt the effectiveness of a search engine.

4.1 Measuring page quality

Given our definition of page quality, the main challenge is how we can measure it. If we want to measure the quality in the strictest sense, we need to contact a large number of web users who visited the page and obtain their feedback on whether they liked the page or not. As previous studies point out, obtaining explicit user feedback is a challenging task and often impractical in real-world settings. Then how can we measure the quality of a page without asking for user feedback?

Our main idea is based on that (1) the creation of a link often indicates that a user likes the page and (2) a high quality page will be liked by most of its visitors, so its popularity may increase more rapidly than others.

First, the success of PageRank shows that when a user creates a link to a page it often indicates that the user is interested in the page. Thus, by observing the existence and creation of the links to a page, we get implicit feedback on the page and can roughly estimate how many people currently like the page.

Second, a high quality page will increase its popularity much more rapidly than others once the page is created, because a large fraction of its visitors will like it when the see it. Therefore, by observing the *increase* (or *time derivative*) of popularity, we may estimate the quality of a page well. Here, we note that the time derivative of popularity can be measured relatively easily. For example, if we use PageRank as the popularity metric of a page, we may download the web multiple times over a period of time, measure how much the PageRank of each page changes over this time period.

The difficult question is exactly how we can use the popularity increase in measuring quality. For instance, is quality directly proportional to popularity increase (i.e., $Q(p) = \frac{d\mathcal{P}(p)}{dt}$ if $\mathcal{P}(p)$ is the popularity of page p)? Should we consider both the current popularity and the popularity increase in measuring quality? Exactly how should we combine these two measures? How do we know whether a particular combination is good?

In order to answer these questions, we take the following approach in this paper: We first assume a simple yet reasonable web-user model that captures the core properties of how users browse web pages. We then analyze this model to derive how the popularity of a page evolves over time. Once we obtain the popularity evolution function, we investigate its property to see how we can estimate quality from popularity evolution. Our analysis will show that we can estimate the quality accurately through the following formula:

$$C \cdot \frac{d\mathcal{P}(p)/dt}{\mathcal{P}(p)} + \mathcal{P}(p), \qquad (1)$$

where C is a constant whose meaning will be clear from our later discussion. The above formula shows that in order to estimate the quality well, we need to consider *both* the current popularity $\mathcal{P}(p)$ and the popularity increase $d\mathcal{P}(p)/dt$. Finally, we evaluate the effectiveness of this quality estimator in an experimental setting. We believe this approach allows us to investigate the problem in a disciplined way and provides a scientific understanding on the core assumptions behind the final ranking mechanism.

In Section 5, we describe the web-user model. In Section 6, we analyze the popularity evolution and obtain the closed form formula for the quality estimator. In Section 8 we present our experimental result.

5. WEB-USER MODEL

We start the description of our web-user model with two definitions of popularity: *(simple) popularity* and *visit popularity*.

Definition 2 (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. □

Under this definition, if 100,000 users out of one million currently like page p_1, its popularity is 0.1.

Notice the subtle difference between the quality of a page and the popularity of a page. The quality is the probability that a web user will like the page *if* the user discovers the page, while the popularity is the *current* fraction of web users who like the page. Thus, a high-quality page may have low popularity because few users are currently aware of the page. While the popularity of a page in its exact sense may be difficult to measure, we may substitute any popularity metric, such as PageRank, as a surrogate to the popularity.

The second notion of popularity, *visit popularity*, measures how many "visits" a page gets in a unit time interval.

Definition 3 (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" the page gets within a unit time interval at time t. □

For example, if 100 users visit page p_1 in the unit time interval from t, and if 200 users visit page p_2 in the same time period, $\mathcal{V}(p_2, t)$ is twice as large as $\mathcal{V}(p_1, t)$.

We also introduce the notion of *user awareness*.

Definition 4 (User awareness) We define the *user awareness* of page p at time t, $\mathcal{A}(p, t)$, as the fraction of web users who are aware of p at time t. □

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. Note that the user awareness of p represents the number of web users who have already visited the page and are aware of it whether they like it or not. In contrast, the popularity of p represents the number of users who know about the page *and* like it.

We assume that a user makes her decision on her liking the page when the user visits the page for the first time and sticks to the decision forever. This assumption is clearly an approximation because some users may arbitrarily change their mind at a later point. However, without any further evidence, it is reasonable to expect that if k users change their positive decision to a negative one, a similar number of users also change their negative decision to a positive one, making the overall number of users in each category the same.

Given the definitions, we can see the following relationship between user awareness, popularity and page quality.

Lemma 1 *The popularity of p at time t, $\mathcal{P}(p, t)$, is equal to the fraction of web users who are aware of p at t, $\mathcal{A}(p, t)$, times the quality of p.*

$$\mathcal{P}(p, t) = \mathcal{A}(p, t) \cdot Q(p) \qquad (2)$$

□

Proof In order for a web user to like the page p, the user has to be aware of p and like the page. The probability that a random web user is aware of the page is $\mathcal{A}(p, t)$ (Definition 4). The probability that the user will like the page is $Q(p)$ (Definition 1). Thus, $\mathcal{P}(p, t) = \mathcal{A}(p, t) \cdot Q(p)$. ■

Note that $\mathcal{P}(p, t)$ and $\mathcal{A}(p, t)$ are functions of time t, but $Q(p)$ is not. That is, we assume that $Q(p)$ is a static value that does not change over time. Therefore, the popularity of page p, $\mathcal{P}(p, t)$, changes over time not because its quality changes, but because users' awareness of the page changes. Later in Section 6.3 we also study the case when the quality $Q(p)$ may also change over time.

We summarize our notation in Table 1.

5.1 Two hypotheses

We now explain two core hypotheses of our model on how users visit web pages. The first hypothesis is based on the random-surfer interpretation of PageRank. In Section 3 we explained that the PageRank of page p is equivalent to the probability that a user will visit the page when the user randomly

Symbol	Meaning
$PR(p)$	PageRank of page p (Section 3)
$Q(p)$	Quality of p (Definition 1)
$\mathcal{P}(p,t)$	(Simple) popularity of p at t (Definition 2)
$\mathcal{V}(p,t)$	Visit popularity of p at t (Definition 3)
$\mathcal{A}(p,t)$	User awareness of p at t (Definition 4)
$\mathcal{I}(p,t)$	Relative popularity increase: $\mathcal{I}(p,t) = \left(\frac{n}{r}\right)\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)}$
r	normalization constant: $\mathcal{V}(p,t) = r\mathcal{P}(p,t)$
n	Total number of web users

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

surfs the web. Given this interpretation, it is reasonable to assume that the number of visitors to a page at time t, $\mathcal{V}(p,t)$, is proportional to its current popularity $\mathcal{P}(p,t)$, which may be measured by PageRank.

Proposition 1 (Popularity-equivalence hypothesis) *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,*

$$\mathcal{V}(p,t) = r\,\mathcal{P}(p,t) \quad (or\ \ \mathcal{V}(p,t) \propto \mathcal{P}(p,t))$$

where r is a normalization constant. □

At an intuitive level, the above hypothesis makes sense because when a page is popular the page is likely to be visited by many people.

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

Proposition 2 (Random-visit hypothesis) *All web users will visit a particular page with equal probability.* □

6. ANALYSIS OF WEB-USER MODEL

We now analyze our web-user model to derive a closed-form formula for the quality estimator. Naively, given Equation 2, we may think that the quality $Q(p)$ can be obtained simply by dividing the current popularity $\mathcal{P}(p,t)$ by the awareness $\mathcal{A}(p,t)$. The problem with this solution is that we do not know the current awareness of a page unless we know the entire history of the page and how many unique users have visited it. Therefore, $Q(p)$ cannot be measured through $\mathcal{P}(p,t)/\mathcal{A}(p,t)$ in most practical settings.

The above discussion brings up an important property desired for the quality estimator: In order to be practical, *the quality estimator should rely only on the quantities that can be measured easily*, such as page popularity. As we briefly described in Section 5, our main idea for quality estimation is that the increase (or time derivative) of popularity may give us a strong hint on the quality of a page. To formally investigate this idea, in Section 6.1 we analyze the popularity evolution of a page under our web-user model. Then in Section 6.2, we take the time derivative of the popularity evolution function and obtain the quality estimator that uses only quantities that are easily measurable.

6.1 Popularity evolution

We now derive the popularity evolution function over time under our model. 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, out of these new users, $Q(p)$ is the fraction that will like the page p, so we can estimate how much its popularity will increase. Therefore, as long as we know the initial popularity of the page p, we can derive its entire popularity evolution over time.

For formal derivation, we first prove the following lemma. The lemma shows that we can learn the *current* user awareness of a page from the history of its *past* popularity. For the proof, we assume that there are n web users in total.

Lemma 2 *The user awareness of p at t, $\mathcal{A}(p,t)$, can be computed from its past popularity through the following formula:*

$$\mathcal{A}(p,t) = 1 - e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt}$$ □

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 $\int_0^t \mathcal{V}(p,t)dt = r\int_0^t \mathcal{P}(p,t)dt$ 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 k times. The probability that the ith 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 $(1-\frac{1}{n})^k$. 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

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

Here we will assume that the number of web users is quite large, so we can approximate the above expression by observing that when $n \to \infty$, $\left(1-\frac{1}{n}\right)^{-n} \to e$. Thus,

$$1-\mathcal{A}(p,t) = e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt} \tag{3}$$

■

Lemma 1 shows that the current popularity of a page can be computed from its current awareness. Lemma 2 shows that the current awareness can be computed from its past popularity. By combining the two lemmas we can compute the current popularity of a page from its past popularity. The following theorem shows the popularity evolution function.

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

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

Here, $\mathcal{P}(p,0)$ is the popularity of the page p at time zero when the page was first created. □

Proof From Lemmas 1 and 2,

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

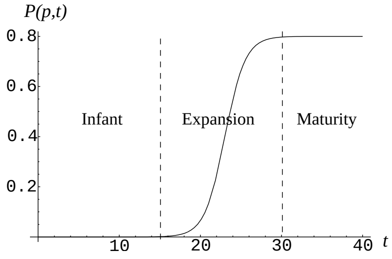

Figure 1: Time evolution of page popularity

If we substitute $e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt}$ with $f(t)$, $\mathcal{P}(p,t)$ is equivalent to $(-\frac{n}{r})(\frac{df}{dt}/f)$. Thus,

$$\left(-\frac{n}{r}\right)\left(\frac{1}{f}\right)\frac{df}{dt} = (1-f)\,Q(p) \tag{4}$$

Equation 4 is known as a Verhulst equation (or logistic growth equation) which often arises in the context of population growth [29]. The solution to the equation is

$$f(t) = \frac{1}{1+Ce^{\frac{r}{n}Q(p)t}}$$

where C is a constant to be determined by the boundary condition. Since $f(t) = e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt}$,

$$e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt} = \frac{1}{1+Ce^{\frac{r}{n}Q(p)t}}. \tag{5}$$

If we take the logarithm of both sides of Equation 5 and differentiate by t,

$$\left(-\frac{r}{n}\right)\mathcal{P}(p,t) = -\frac{\left(\frac{r}{n}\right)Q(p)\,C\,e^{\frac{r}{n}Q(p)t}}{1+Ce^{\frac{r}{n}Q(p)t}}.$$

After rearrangement, we get

$$\mathcal{P}(p,t) = \frac{CQ(p)}{C+e^{-\frac{r}{n}Q(p)t}}. \tag{6}$$

We now determine the constant C. From Equation 6

$$\mathcal{P}(p,0) = \frac{CQ(p)}{C+1}. \tag{7}$$

Thus,

$$C = \frac{\mathcal{P}(p,0)}{Q(p)-\mathcal{P}(p,0)} \tag{8}$$

After rearrangement, we finally get

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

Based on the result of the above theorem, we show an example of the popularity evolution of a page in Figure 1. We assume $Q(p) = 0.8$, $n = 10^8$, $r = 10^8$ and $\mathcal{P}(p,0) = 10^{-8}$. Roughly, these parameters correspond to the case where there are 100 million web users and only one user liked the page p at its creation. The quality is relatively high at 0.8. The horizontal axis corresponds to the time. The vertical axis corresponds to the popularity $\mathcal{P}(p,t)$ at the given time.

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 $t = 15$) the page is barely noticed by web users and has practically zero popularity. At some point ($t = 15$), however, the page enters the second expansion stage ($t = 15$ and 30), where the popularity of the page suddenly increases. In the third maturity stage, the popularity of the page stabilizes at a certain value. Note that this "sigmoidal" evolution of popularity has been experimentally observed in the site popularity-evolution data collected by web tracking companies (e.g., NetRatings [18]).

We also note that the eventual popularity of p is equal to its quality value 0.8. The following corollary shows that this equality holds in general.

Corollary 1 *The popularity of page p, $\mathcal{P}(p,t)$, eventually converges to $Q(p)$. That is, when $t \to \infty$, $\mathcal{P}(p,t) \to Q(p)$.* □

Proof From Theorem 1,

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

When $t \to \infty$, $e^{-\left[\frac{r}{n}Q(p)\right]t} \to 0$. Thus,

$$\begin{aligned}\mathcal{P}(p,t) &= \frac{\mathcal{A}(p,0)\;Q(p)}{\mathcal{A}(p,0)+[1-\mathcal{A}(p,0)]\;e^{-\left[\frac{r}{n}Q(p)\right]t}}\\ &\to \frac{\mathcal{A}(p,0)\;Q(p)}{\mathcal{A}(p,0)} = Q(p).\end{aligned} \qquad \blacksquare$$

The result of this corollary is reasonable. When all users are aware of the page, the fraction of all web users who like the page is the quality of the page.

The result of Figure 1 confirms our earlier assertion that the popularity of a page is not a good estimator of its quality for new pages: During the infant and the expansion stage ($t < 30$), the popularity of the page is significantly lower than its true quality value. It is only in the maturity stage when the popularity reflects the true quality of the page.

Finally, we note that the popularity evolution in Figure 1 is monotone since we assume that the quality of a page is a static value that does not change. The quality, however, may also change over time, for example, when the page is updated or when many high-quality pages appear on the web and the users' expectation on the page gets higher. In Section 6.3 we extend our model and study the case when the quality of a page changes. For now, we assume that the quality value of a page is static.

6.2 Quality estimator

We can analyze the popularity evolution function derived in the previous section to explore whether its time derivative can be used to estimate the quality of a page. The following lemma provides the core relationship between the time derivative of page popularity and page quality.

Lemma 3 *The quality of a page is proportional to its popularity increase and inversely proportional to its current popularity. It is also inversely proportional to the fraction of the users who are unaware of the page, $1-\mathcal{A}(p,t)$.*

$$Q(p) = \left(\frac{n}{r}\right)\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)\;(1-\mathcal{A}(p,t))} \tag{9}$$

□

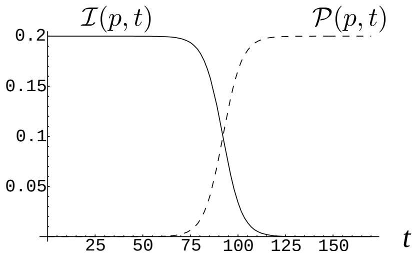

Figure 2: Time evolution of $\mathcal{I}(p,t)$ and $\mathcal{P}(p,t)$ as predicted by the model.

Proof By differentiating the equation in Lemma 1, we get

$$\frac{d\mathcal{P}}{dt} = \frac{d\mathcal{A}}{dt} Q(p). \tag{10}$$

From Lemma 2,

$$\begin{aligned}\frac{d\mathcal{A}}{dt} &= -\frac{d}{dt} e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt} \\ &= -\left(e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt}\right)\left(-\frac{r}{n}\mathcal{P}(p,t)\right) \\ &= (1-\mathcal{A}(p,t))\left(\frac{r}{n}\mathcal{P}(p,t)\right). \end{aligned} \tag{11}$$

From Equations 10 and 11, we get

$$Q(p) = \left(\frac{n}{r}\right)\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)\ (1-\mathcal{A}(p,t))}. \quad ■$$

In Equation 9, note that two main factors, $d\mathcal{P}(p,t)/dt$ and $\mathcal{P}(p,t)$, are measurable in practice by downloading the web multiple times while $1-\mathcal{A}(p,t)$ cannot be easily measured. Therefore, for now, we ignore the unmeasurable factor $1 - \mathcal{A}(p,t)$ from the equation and study the property of the remaining factors $\left(\frac{n}{r}\right)\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)}$ as the quality estimator. Intuitively, $d\mathcal{P}(p,t)/dt$ is the popularity increase of the page and $\mathcal{P}(p,t)$ is the current popularity, so the ratio $\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)}$ is the relative popularity increase of the page. For convenience, we use the symbol $\mathcal{I}(p,t)$ to represent $\left(\frac{n}{r}\right)\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)}$ and refer to it as the *relative popularity increase function*.

In Figure 2, we show the time evolution of $\mathcal{I}(p,t)$ when $Q(p) = 0.2$, $n = 10^8$, $r = 10^8$, and $\mathcal{P}(p,0) = 10^{-9}$. The horizontal axis is the time and the vertical axis shows the value of the function. The solid line in the graph shows the relative popularity increase $\mathcal{I}(p,t)$. We also show the time evolution of the popularity $\mathcal{P}(p,t)$ as a dashed line in the figure for the comparison purpose. We obtained these graphs analytically using the equation of Theorem 1.

From the graph, we can see that the relative popularity increase, $\mathcal{I}(p,t)$, is an excellent estimator for the page quality $Q(p)$ when the page has just been created ($t < 70$). During this time, $\mathcal{I}(p,t) \approx 0.2 = Q(p)$. As time goes on, however, $\mathcal{I}(p,t)$ loses its merit as the estimator of $Q(p)$: $\mathcal{I}(p,t)$ gets much smaller than $Q(p)$ for $t > 120$. Fortunately, when $\mathcal{I}(p,t)$ is not a good quality estimator, we can see that $\mathcal{P}(p,t)$ is a very good estimator of $Q(p)$ ($t > 120$). That is, $\mathcal{I}(p,t)$ and $\mathcal{P}(p,t)$ are *complementary* to each other as the quality estimator.

Intuitively, the relative effectiveness of $\mathcal{P}(p,t)$ and $\mathcal{I}(p,t)$ as the quality estimator makes sense. When a page has just been created, most users are unaware of the page, so its popularity $\mathcal{P}(p,t)$ does not reflect its quality well. However, the users who visit the page are mostly first-time visitors, so if the page is of high quality, its popularity will increase very rapidly, making the relative popularity increase $\mathcal{I}(p,t)$ a good quality estimator. As time goes on, however, most users get aware of the page, so the popularity of the page cannot increase any further. Fortunately at this point, the fraction of web users who like the page $\mathcal{P}(p,t)$ is equivalent to its quality, making it a good quality estimator.

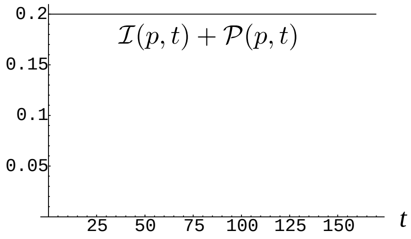

Figure 3: Time evolution of $\mathcal{I}(p,t)+\mathcal{P}(p,t)$.

From the shape of the two curves in Figure 2 we can expect that we may estimate the quality of the page accurately if we add these two functions. In Figure 3, we show the time evolution of this addition, $\mathcal{I}(p,t)+\mathcal{P}(p,t)$, for the same parameters as in Figure 2. We can see that $\mathcal{I}(p,t)+\mathcal{P}(p,t)$ is a straight line at the quality value 0.2. The following theorem generalizes this observation and shows that $\mathcal{I}(p,t)+\mathcal{P}(p,t)$ is indeed an accurate quality estimator.

Theorem 2 *The quality of page p, $Q(p)$, is always equal to the sum of its relative popularity increase $\mathcal{I}(p,t)$ and its popularity $\mathcal{P}(p,t)$.*

$$Q(p) = \mathcal{I}(p,t) + \mathcal{P}(p,t) \quad □$$

Proof We first restate Equation 11:

$$\frac{d\mathcal{A}(p,t)}{dt} = (1-\mathcal{A}(p,t))\left(\frac{r}{n}\right)\mathcal{P}(p,t)$$

If we multiply the above equation by $Q(p)$, we get

$$Q(p)\frac{d\mathcal{A}(p,t)}{dt} = (Q(p) - Q(p)\mathcal{A}(p,t))\left(\frac{r}{n}\right)\mathcal{P}(p,t),$$

which can be simplified to

$$\frac{d\mathcal{P}(p,t)}{dt} = (Q(p) - \mathcal{P}(p,t))\left(\frac{r}{n}\right)\mathcal{P}(p,t).$$

If we divide the equation by $\frac{r}{n}\mathcal{P}(p,t)$ and add $\mathcal{P}(p,t)$ to both sides, we get

$$\frac{d\mathcal{P}(p,t)/dt}{(r/n)\mathcal{P}(p,t)} + \mathcal{P}(p,t) = Q(p) \quad ■$$

The above theorem shows that under our web user model we can compute the quality of a page by measuring its relative popularity increase and current popularity. Based on this result, we define $\mathcal{I}(p,t)+\mathcal{P}(p,t)$ as the *quality estimator* of p, $\hat{\mathcal{Q}}(p,t)$:

$$\begin{aligned}\hat{Q}(p,t) &= \mathcal{I}(p,t) + \mathcal{P}(p,t) \\ &= \left(\frac{n}{r}\right)\left(\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)}\right) + \mathcal{P}(p,t)\end{aligned} \tag{12}$$

Later in Section 8, we evaluate the effectiveness of the above quality estimator experimentally.

6.3 Changing quality

So far we have assumed that the quality $Q(p)$ of a page is a constant that does not change over time. In this section, we analyze the scenario where the quality also changes. Our main goal is to understand how our quality estimator should be updated to handle this scenario. Before we describe our formal analysis and result, we use a simple example to illustrate our main finding.

Example 1 We assume that the page p was originally of quality $Q_1 = 0.8$ from $t = 0$ until $t = 30$. At $t = 30$, the quality suddenly drops to $Q_2 = 0.4$. We show the popularity evolution under this scenario in Figure 4. The graph was obtained analytically based on the same parameters as in Section 6.1.

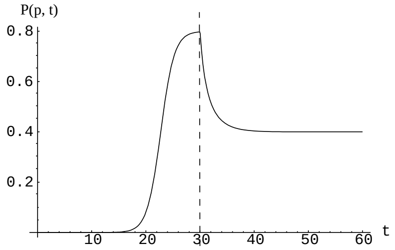

Figure 4: Popularity evolution when quality drops at $t = 30$

As before, the popularity increases from $t = 0$ until $t = 30$ as more people visit the page and get aware of it. By $t = 30$, the popularity becomes close to its quality value 0.8. After $t = 30$, however, the popularity gradually decreases due to the drop in quality: the people who visit the page again after $t = 30$ realize that its quality is not as good as it used to be and many of them stop liking the page. This decrease continues until the popularity stabilizes at the new quality value 0.4. □

In the above example, we note that the situation between $t = 0$ and $t = 30$ is essentially the same as in the previous section, so $\left(\frac{n}{r}\right) \frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)} + \mathcal{P}(p,t)$ is a good quality estimator during this time. But what will be a good quality estimator after $t = 30$? Does our estimator still work well in this region and estimate the correct quality value Q_2 after $t = 30$? Our analysis shows that our estimator is still valid even in this region. That is, $\left(\frac{n}{r}\right) \frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)} + \mathcal{P}(p,t) = Q_2$ for $t > 30$. In general, we can prove the following theorem that shows that our estimator is still valid even after a quality change:

Theorem 3 *We assume that the quality of page p, $Q(p)$, changes from Q_1 to Q_2 at time T. Then*

$$Q_2 = \left(\frac{n}{r}\right) \frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)} + \mathcal{P}(p,t) \quad \textit{for} \quad t > T \tag{13}$$

□

Proof After time T, we can put users into three groups: (1) the users who visited the page before T (2) the users who visited the page after T and (3) the users who never visited the page. Of course, some users may belong to both groups (1) and (2) if they visited the page before *and* after T. We use the the notation u_1 and u_2 to represent the group (1) and (2), respectively. In Figure 5, we show the Venn diagram showing u_1, u_2 and the total user group U. We use $|u_1|$ and $|u_2|$ to represent the relative size of each group: i.e., the fraction of users who belong to u_1 and u_2, respectively.

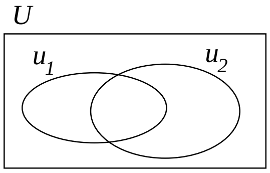

Figure 5: Venn diagram for user groups after time T

At time $t > T$, we note that the users who belong to u_2 have seen the page when its quality is Q_2, so Q_2 fraction of the users in u_2 end up liking the page. The users who belong to $(u_1 - u_2)$ group (the users who visited the page before T but not after T) still believes that the quality of the page is Q_1 because they haven't seen the new page. Therefore, out of $|u_1 - u_2|$ users, Q_1 fraction still like the page. Of course, the users who are not in either u_1 or u_2 cannot like the page because they have never visited the page. Overall, the fraction of web users who like the page at time $t > T$ is

$$\mathcal{P}(p,t) = Q_1|u_1 - u_2| + Q_2|u_2|.$$

Note that after $t > T$, the u_1 group remains the same but the u_2 group gradually expands as more people visit the page after T. We denote this time dependence using the notation $u_2(t)$ for u_2 but simply u_1 for u_1. Then,

$$\mathcal{P}(p,t) = Q_1|u_1 - u_2(t)| + Q_2|u_2(t)| \quad \text{for } t > T. \tag{14}$$

We now compute $|u_2(t)|$ and $|u_1 - u_2(t)|$ in order to compute $\mathcal{P}(p,t)$. $|u_2(t)|$ is the fraction of users who visit p from time T to t. From the proof of Lemma 2, it is easy to see that this fraction is given by

$$|u_2(t)| = 1 - e^{-\frac{r}{n}\int_T^t \mathcal{P}(p,t)dt}. \tag{15}$$

In computing $|u_1 - u_2(t)|$, we note that the user groups u_1 and u_2 are independent. That is, according to our random-visit hypothesis, the probability that a user visits the page p at t is independent of his past visit history, so whether the user visits p after time T is independent of whether he visited the page before T. Given this independence, the size of the intersection $u_1 \cap u_2(t)$ can be computed by simple multiplication

$$|u_1 \cap u_2(t)| = |u_1| \cdot |u_2(t)|.$$

Then

$$|u_1 - u_2(t)| = |u_1| - |u_1 \cap u_2(t)| = |u_1| - |u_1||u_2(t)|$$

and

$$\begin{aligned}\mathcal{P}(p,t) &= Q_1|u_1 - u_2(t)| + Q_2|u_2(t)| \\ &= Q_1|u_1| - Q_1|u_1||u_2(t)| + Q_2|u_2(t)| \\ &= Q_1|u_1| + (Q_2 - Q_1|u_1|)|u_2(t)|.\end{aligned}$$

We now differentiate this equation by t and get

$$\frac{d\mathcal{P}(p,t)}{dt} = (Q_2 - Q_1|u_1|)\frac{d|u_2(t)|}{dt}. \tag{16}$$

From Equation 15, we know that

$$\begin{aligned}\frac{d|u_2(t)|}{dt} &= \frac{r}{n}\mathcal{P}(p,t)e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt} \\ &= \frac{r}{n}\mathcal{P}(p,t)(1-|u_2(t)|).\end{aligned}$$

Therefore, Equation 16 becomes

$$\begin{aligned}\frac{d\mathcal{P}(p,t)}{dt} &= (Q_2 - Q_1|u_1|)\frac{r}{n}\mathcal{P}(p,t)(1-|u_2(t)|) \\ &= \frac{r}{n}\mathcal{P}(p,t)[Q_2 - \{Q_1|u_1| - Q_1|u_1||u_2(t)| \\ &\qquad + Q_2|u_2(t)|\}] \\ &= \frac{r}{n}\mathcal{P}(p,t)[Q_2 - \{Q_1|u_1 - u_2(t)| + Q_2|u_2(t)|\}] \\ &= \frac{r}{n}\mathcal{P}(p,t)[Q_2 - \mathcal{P}(p,t)] \quad \text{(from Eq. 14)}\end{aligned}$$

If we divide the above equation by $\frac{r}{n}\mathcal{P}(p,t)$ and add $\mathcal{P}(p,t)$, we get

$$\left(\frac{n}{r}\right)\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)} + \mathcal{P}(p,t) = Q_2. \qquad \blacksquare$$

7. MEASURING QUALITY FROM WEB SNAPSHOTS

In the previous sections we discussed how we can estimate the quality of a page on the basis of its present popularity and its instantaneous time derivative. In practice, however, the time derivative cannot be measured instantaneously, but only can be approximated through the increase of PageRank at *discrete* time points. That is, we take the snapshots of the web at times $t_1, t_2, t_3, \ldots$, compute PageRank of pages from each snapshot and approximate Equation 12 with

$$\hat{Q}(p,t_i) = \frac{n}{r}\left[\frac{\Delta PR(p,t_i)/\Delta t_i}{PR(p,t_i)}\right] + PR(p,t_i) \qquad (17)$$

where, $PR(p,t_i)$ is the PageRank of p at t_i, $\Delta PR(p,t_i) = PR(p,t_i) - PR(p,t_{i-1})$, and $\Delta t_i = t_i - t_{i-1}$.

Unfortunately, this discrete measurement may lead to an error for the following reasons.

1. *Approximation error*: $\Delta PR(p)/\Delta t$ is an approximation of $dPR(p)/dt$. Thus, the values can be different, particularly when Δt is large.

2. *Quality change during measurement*: In our theoretical derivations, we assumed that the quality remains constant during measurement.[1] This assumption is reasonable when we can measure the derivative instantaneously, but when it is measured over a time period, it is possible that the quality may change during the time.

3. *Time lag*: Consider the time period after t_2 but before t_3. Since we haven't captured the t_3 snapshot, the most recent quality estimate is the one computed from the t_1 and t_2 snapshots. That is, during the interval (t_2, t_3), we use the quality value measured in (t_1, t_2). This *time lag* between the quality measurement and its use may lead to an error if the quality changes after t_2 even if it did not change during (t_1, t_2).

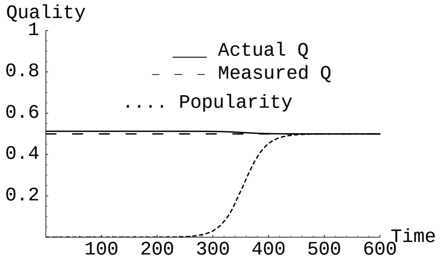

Figure 6: True and estimated quality values for a static quality scenario

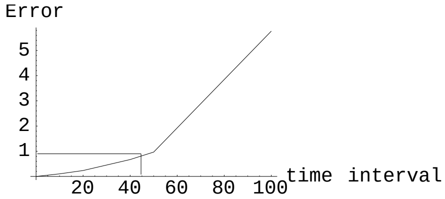

Figure 7: Error values as we increase the time interval between two consecutive popularity measurements

To investigate the the impact of these errors, we use the following three scenarios.

1. *Static quality*: We consider a page p whose quality value Q remains constant at 0.5. We assume that we take the snapshot of the web every $t = 1$ time unit, and we recompute the quality value using the most two recent snapshots. That is, at $t = i$, we estimate $\hat{Q}$ from the popularity values at $t = i$ and $t = i - 1$. We use the quality value estimated at $t = i$ during $t \in (i, i+1)$. Figure 6 shows the true quality Q, the popularity $\mathcal{P}$, and the estimated quality $\hat{Q}$ over time under this setting.

   In this scenario, the only source of error is the approximation error, because the quality value remains the same all the time. The figure shows that this error is negligible in this scenario; Q and $\hat{Q}$ are almost identical. We also see that our quality estimator $\hat{Q}$ works well as a "eventual popularity predictor." That is, at any time point, $\hat{Q}$ gives the value 0.5, which is the same as the eventual popularity of the page. In contrast, the current popularity $\mathcal{P}$ is not a good predictor of the eventual popularity; From $t = 1$ until $t = 400$, $\mathcal{P}$ is significantly smaller than 0.5.

   The magnitude of the approximation error will clearly depend on the length of interval Δt. In order to study this impact, we repeat the same experiment for different Δt values and show the result in Figure 7. The horizontal axis is Δt and the vertical axis is the error, $|Q - \hat{Q}|$, for the given Δt. As we expect, the error becomes larger at Δt grows. For example, when $\Delta t = 45$ (about 10% of the time it took for the page to obtain the eventual popularity), the error is 0.33 (67% relative error).

2. *Slow change in quality*: We consider a page p whose ini-

[1] Note that our result in Section 6.3 shows that a quality change *before* or *after* measurement does not affect the validity of our estimator. However, our analysis does not guarantee its correctness if there is a change *during* measurement.

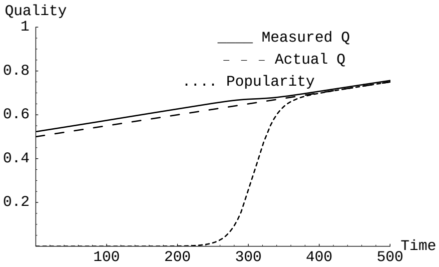

Figure 8: Graph of actual and measured quality values

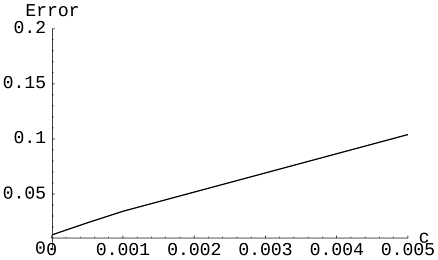

Figure 9: Graph showing error values as we increase the rate of change in quality

tial quality Q is 0.4. The quality value increases slowly over time according to the relation $Q(t) = 0.4{+}0.0006t$, reaching 0.7 at $t = 500$. We measure the quality estimate $\hat{Q}$ after every unit time interval. We plot the true quality, the popularity, and the estimated quality values in Figure 8 under this scenario.

From the figure, we observe the following:

(a) Our estimator $\hat{Q}$ still measures the true quality well; At every time point, $\hat{Q} \approx Q$.

(b) $\hat{Q}$ is not a good predictor of the eventual popularity. For example $\hat{Q} \approx 0.4$ at $t = 1$, but the eventual popularity is 0.7 at $t = 500$ in this graph. It should be noted, however, that $\hat{Q}$ is a better predictor of the eventual popularity than the current popularity $\mathcal{P}$. For example, $\hat{Q} \approx 0.4$ at $t = 1$, which is much closer to the eventual popularity 0.7 than $\mathcal{P} \approx 0$ at $t = 1$.

The magnitude of the error in $\hat{Q}$ may depend on the rate of change of the quality in this scenario. To investigate this issue, we repeat the same analysis using $Q(t) = 0.5{+}ct$ for multiple values of c. Figure 9 shows the result, where the horizontal axis is c, the change rate, and the vertical axis is the error $|Q - \hat{Q}|$ at the given c. For example, when $c = 0.005$ (about 1% increase in quality in one time unit), $|Q - \hat{Q}| \approx 0.1$ (20% relative error in the quality estimation). As expected, with increase in the value of c, the error also increases.

3. *Rapid change in quality*: Finally, we consider a scenario where the quality Q of the page rapidly fluctuates over time. As we show in Figure 10, the quality of the page changes according to a sinusoidal relation. We assume that we measure the quality of the page after every unit time interval.

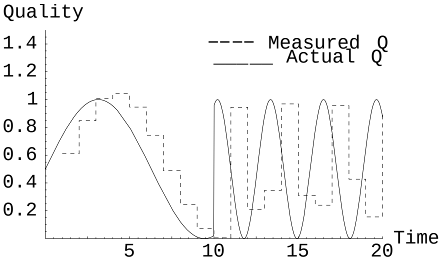

Figure 10: Graph of actual and measured quality values

Overall, we can see that the overall shapes of Q and $\hat{Q}$ graphs are similar. However, the time lag becomes an important source of error in this scenario; The $\hat{Q}$ curve is one time unit behind the Q curve, which makes the two values very different at many time points. Also, $\hat{Q}$ is a very crude approximation of Q; Even if we ignore the time-lag error, $\hat{Q}$ and Q values are significantly different sometimes due to the jagged nature of $\hat{Q}$. Finally, we note that $\hat{Q}$ is not a good predictor of the eventual popularity; Because Q fluctuates frequently, there is no correlation between the current quality estimate $\hat{Q}$ and the eventual popularity. In summary, $\hat{Q}$ is not very effective when the quality rapidly changes over time.

8. EXPERIMENTS

Given that our ultimate goal is to find high-quality pages and rank them highly in search results, the best way to evaluate our quality estimator is to implement it on a search engine and see how well users perceive our new ranking. Before we embark on this enormous endeavor, we wanted to check the potential of our proposed quality estimator in a more practical and manageable setting.

Evaluating a web ranking metric is a challenging task because of its subjectivity and the lack of standard corpus. The relevance and quality of a page is clearly a subjective notion, so the best way of measuring the effectiveness of a ranking metric is to ask a large number of users to go over a collection of web pages carefully and provide their feedback on the perceived quality of each page. This task is clearly time consuming and expensive. Recognizing this challenge, the IR community has collaboratively constructed a standard evaluation corpus, called TREC [27], which also includes a special subcollection of web documents. Unfortunately, this dataset is not well suited for our evaluation, because (1) it only contains a single snapshot of the web, making it impossible to measure the evolution of PageRank and (2) the dataset indicates only the binary relevance (either 0 or 1) of each page to a number of predefined queries. With the binary relevance, we cannot *rank* the pages based on their quality and compare this ranking to the one from our quality metric.

Thus, we take an alternative approach to evaluating the potential of our quality estimator. Our main idea for evaluation is that when the quality value does not change significantly over time, the popularity of a page eventually converges to its quality. That is, the eventual popularity of a page is a good estimator of its quality. Thus, for the pages with stable quality, if we can wait long enough, our estimated quality should a good "predictor" of the eventual PageRank. Based on this idea, we capture multiple snapshots of the web, compute page quality, and compare today's quality value with the PageRank value in the future. Admittedly, this evaluation is not perfect because the quality is compared against future PageRank, a metric that it tries to replace. However, with the lack of the true quality value for each page we believe that this comparison, at the very least, will show the potential of our estimator.

8.1 Description of dataset

Due to our limited network and storage resources, we had to restrict our experiments to a relatively small subset of the web. In our experiment we downloaded pages on 154 web sites (e.g., `acm.org`, `hp.com`, etc.) four times over the period of six months. The list of the web sites were collected from the Open Directory (`http://dmoz.org`). The timeline of our snapshots is shown in Figure 11. Roughly, the first three snapshots were taken with one-month intervals between them and the last snapshot was taken four months after the third snapshot. We refer to the times of the four snapshots as t_1, t_2, t_3 and t_4. Later in this section, we will use the first three snapshots to compute the quality of pages and evaluate how well the earlier quality values "predict" the "future" PageRanks at t_4.

Our snapshots were quite complete mirrors of the 154 web sites. We downloaded pages from each site until we could not reach any more pages from the site or we downloaded the maximum of 200,000 pages. Out of 154 web sites, only four web sites had more than 200,000 pages. The number of pages that we downloaded in each snapshot ranged between 4.6 million pages and 5 million pages. Since we were interested in comparing our estimated page quality with the future PageRank, we first identified the set of pages downloaded in all snapshots. Out of 5 million pages, 2.7 million pages were common in all four snapshots. We then computed the PageRank values from the subgraph of the web obtained from these 2.7 million pages for each snapshot.

8.2 Quality and future PageRank

Using the dataset just described, we now investigate how well our quality estimator predicts the future PageRank.

Stability of quality. In Section 7 we showed that our quality estimator is a good predictor of the future PageRank only when the quality does not change significantly over time.[2] We first investigate how many pages in our snapshots have stable quality values.

To check the stability, we compute *three* quality values, $\hat{Q}_2$, $\hat{Q}_3$ and $\hat{Q}_4$, from our four snapshots, where $\hat{Q}_i$ is measured based on the PageRank values of the t_i and t_{i-1} snapshots using Equation 17 assuming 0.1 for r/n. (The choice of 0.1 is explained later). In Figure 12 we show the histogram of the relative quality difference between $\hat{Q}_2$ and $\hat{Q}_3$ (white bars) and between $\hat{Q}_3$ and $\hat{Q}_4$ (gray bars). From the histogram we can see that the vast majority of pages in our snapshots have stable quality. For example, the first white and gray bars indicate that more than 99% pages show less than 10% relative difference between $\hat{Q}_2$ and $\hat{Q}_3$ and about 90% pages show less than 10% difference between $\hat{Q}_3$ and $\hat{Q}_4$.

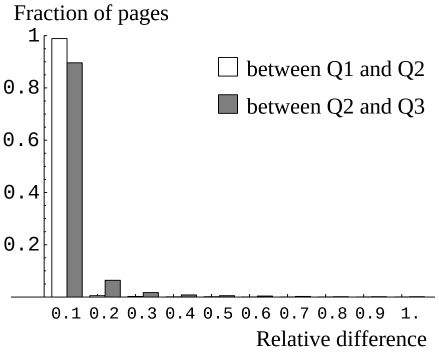

Figure 12: Relative quality difference between snapshots

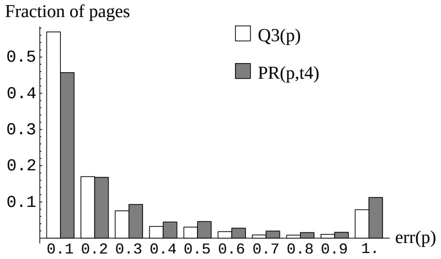

Figure 13: Histogram of relative errors

Prediction accuracy. We now compare the prediction accuracy of the "future" PageRank $PR(p, t_4)$ when we use the "current" PageRank $PR(p, t_3)$ or our quality estimator $\hat{Q}_3(p)$ as the $PR(p, t_4)$ predictor. For the comparison, we compute the following average relative "error":

$$err(p) = \begin{cases} \left| \frac{PR(p,t_4) - \hat{Q}_3(p)}{PR(p,t_4)} \right| & \text{for } \hat{Q}_3(p) \\ \\ \left| \frac{PR(p,t_4) - PR(p,t_3)}{PR(p,t_4)} \right| & \text{for } PR(p,t_3) \end{cases}$$

From this comparison, we observe that the average error is 0.45 for $\hat{Q}_3(p)$ while it is 0.74 for $PR(p, t_3)$. That is, our quality estimator $\hat{Q}_3(p)$ shows about 39% more accuracy in predicting the future PageRank on average.[3] Assuming that the PageRank at t_4 is closer to the true quality of pages, this result strongly indicates that our estimator measures the quality much more accurately than the current PageRank.

In Figure 13, we report more detailed result from this comparison. In the graph, we show the distribution of the relative errors for $\hat{Q}_3(p)$ and $PR(p, t_3)$. The white bars correspond to the histogram of $\hat{Q}_3(p)$ and the gray bars correspond to $PR(p, t_3)$. For example, from the first bars of the graph we can see that $\hat{Q}_3(p)$ shows less than 0.1 relative error for 56% pages, while $PR(p, t_3)$ shows similar error for 45% of the pages. When the relative error is larger than 1, we put

[2] This statement does not mean that our quality estimator is not useful when there are quality changes. Our estimator still measures the true quality value well even with quality changes, but the true quality value may not be the same as the future PageRank.

[3] To show their difference more clearly we compute the average error only for the pages for which $\hat{Q}_3(p)$ and $PR(p, t_3)$ give more than 5% different prediction for the future PageRank.

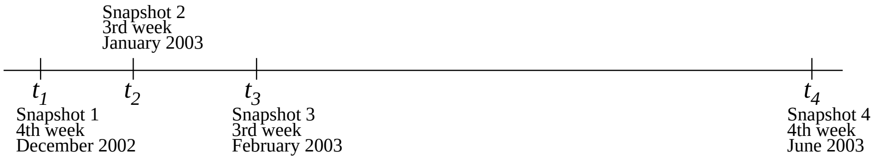

Figure 11: The timeline that our four snapshots were taken

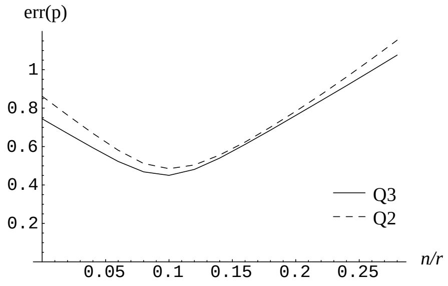

Figure 14: Choice of n/r

them into the last bin labeled as 1. This graph shows that our quality estimator $\hat{Q}_3(p)$ leads to smaller errors for more pages than $PR(p, t_3)$. We also conducted similar comparison using $\hat{Q}_2(p)$ as the quality estimator and obtained comparable results.

Estimation of n/r. We now explain our choice of 0.1 for the parameter n/r. In our theoretical model, n corresponds to the total number of web users and r is the normalization constant in our popularity-equivalence hypothesis. In practice, n/r determines how much "weight" we assigns to the relative popularity increase term. For example, when $n/r = 0$, our quality estimator reduces to $\hat{Q}(p, t) = \mathcal{P}(p, t)$, which completely ignore the popularity increase in estimating quality. That is, when r/n is small, our estimator becomes closer to the traditional PageRank metric and gets more "conservative" in using the popularity increase.

To determine the best choice for n/r, we use the following approach: We measure the relative error between $\hat{Q}(p)$ and $PR(p, t_4)$ for multiple n/r values and we pick the value that leads to the smallest difference. In Figure 14, we show the error between $\hat{Q}_2$ and $PR(t_4)$ and between $\hat{Q}_3$ and $PR(t_4)$ for multiple n/r values. From the graph, we can see that we get minimal error around $n/r = 0.1$ both for Q_2 and Q_3. Based on this result, we use $n/r = 0.1$ for other experiments in this section. We emphasize, however, that our quality estimator is still as good as or better than the current PageRank as the future PageRank predictor when we use any value between 0 and 0.1 for n/r: As n/r gradually decreases from 0.1 to 0, our estimator becomes closer to PageRank and error gets closer to that of PageRank.

9. CONCLUSION

In this paper, we investigated the problem of page quality, including how to quantify the subjective notion of page quality, how well existing search engines measure the quality, and how we might measure the quality of a page more directly. In our study, we proposed a reasonable definition for page quality and we derived a practical way of estimating the quality based on a careful analysis of a reasonable web-user model. Finally, we evaluated the potential of our quality estimator through an experiment.

At a very high level, we may consider our proposed quality estimator as a third-generation ranking metric. The first-generation ranking metric (before PageRank) judged the relevance and quality of a page mainly based on the content of a page without much consideration of web link structure. Then researchers [16, 21] proposed second-generation ranking metrics that exploited the link structure of the web. In our study, we argued that we can further improve the ranking metrics by considering not just the current link structure, but also the *evolution* and *change* in the link structure.

As more digital information becomes available, and as the web further matures, it will get increasingly difficult for new pages to be discovered by users and get the attention that they deserve. We believe that our new ranking metric will help us alleviate this "information imbalance" problem that only established pages are repeatedly looked at by users. Our metric can identity these high-quality pages much earlier than existing metrics and shorten the time it takes for new pages to get noticed.

9.1 Discussion and future work

While our result indicates that our quality metric is a good way to measure the quality of a page in practice, we discuss some of the limitations of our work and potential venues for future work.

- *Statistical Noise*: One potential problem with the quality metric is that it may be adversely affected by noise for pages with very low popularity. When we are measuring the rare event of a page with low popularity receiving a new link, there is the potential that noise could cause such a page to be promoted prematurely. Further work is required to investigate how best to smooth out the curve, including perhaps adjusting the web download intervals depending on the current PageRank values. For example, for low-PageRank pages, we may want to compute the PageRank increase over a longer period than high-PageRank pages in order reduce the impact of noise.

- *Scale of the data*: Our experiment was based on a small subset of the web. While our result indicated improvement over the PageRank metric, it will be interesting to see how well our quality estimator works for a larger dataset.

- *Application to web traffic data*: While in this paper we used the web link structure and its evolution to measure popularity (and thus quality), our estimator can be similarly applied to the web traffic data. That is, assuming that the visit popularity is equivalent to the (simple) popularity (Proposition 1), if we can measure how many people visit a particular web site and how quickly the number of visits increases over time, we can use our quality estimator to measure the quality of the site based on this traffic data. It will be interesting to see how this traffic-based quality estimate is different from our link-based quality estimate and which quality estimate users prefer.

10. ACKNOWLEDGEMENTS

We thank our anonymous reviewers for providing thoughtful comments to improve the paper. In particular, the current proof for Theorem 2 is partially based on the idea suggested by a reviewer. We also thank our "shepherd" Frank Tompa for providing the guidance through the final revision process.

This material is based upon work supported partially by the National Science Foundation under Grant No. 0347993. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

11. REFERENCES

[1] S. Abiteboul, M. Preda, and G. Cobna. Adaptive on-line page importance computation. In *Proceedings of the International World-Wide Web Conference*, May 2003.

[2] D. Achlioptas, A. Fiat, A. R. Karlin, and F. McSherry. Web search via hub synthesis. In *IEEE Symposium on Foundations of Computer Science*, pages 500–509, 2001.

[3] R. Albert, A.-L. Barabasi, and H. Jeong. Diameter of the World Wide Web. *Nature*, 401(6749):130–131, September 1999.

[4] R. Baeza-Yates, F. Saint-Jean, and C. Castillo. Web dynamics, age and page quality. In *Proceedings of SPIRE 2002*, September 2002.

[5] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: authority-based keyword search in databases. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, August 2004.

[6] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[7] 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.

[8] J. Cho and S. Roy. Impact of search engines on page popularity. In *Proceedings of the International World-Wide Web Conference*, May 2004.

[9] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[10] F. Geerts, H. Mannila, and E. Terzi. Relational link-based ranking. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, August 2004.

[11] R. Goldman, N. Shivakumar, S. Venkatasubramanian, and H. Garcia-Molina. Proximity search in databases. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 26–37, 1998.

[12] S. P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49, December 1996.

[13] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the International World-Wide Web Conference*, May 2002.

[14] S. Kamvar, T. Haveliwala, and G. Golub. Adaptive methods for the computation of pagerank. In *Proceedings of International Conference on the Numerical Solution of Markov Chains*, September 2003.

[15] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the International World-Wide Web Conference*, May 2003.

[16] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.

[17] S. Mizzaro. Measuring the agreement among relevance judges. In *Proceedings of MIRA Conference*, April 1999.

[18] Nielsen NetRatings. `http://www.nielsen-netratings.com/`.

[19] Npd search and portal site study. Available at `http://www.npd.com/press/releases/press_000919.htm`.

[20] S. Olsen. Does search engine's power threaten web's independence? Available at `http://news.com.com/2009-1023-963618.html`, October 2002.

[21] 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`.

[22] 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.

[23] S. E. Robertson and K. Sparck-Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1975.

[24] G. Salton. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall Inc., 1971.

[25] G. Salton and M. J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, 1983.

[26] 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.

[27] TREC: Text retrieval conference. `http://trec.nist.gov`.

[28] A. C. Tsoi, G. Morini, F. Scarselli, M. Hagenbuchner, and M. Maggini. Adaptive ranking of web pages. In *Proceedings of the International World-Wide Web Conference*, May 2003.

[29] F. Verhulst. *Nonlinear Differential Equations and Dynamical Systems*. Springer Verlag, 2nd edition, 1997.

[30] Y. Wang and D. DeWitt. Computing pagerank in a distributed internet search system. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, August 2004.

[31] S. Wartick. Boolean operations. *Information Retrieval: Data Structures and Algorithms*, pages 264–292, 1992.