Page Quality: In Search of an Unbiased Web Ranking

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

Robert E. Adams
UCLA Computer Science
Los Angeles, CA 90095
robadams@cs.ucla.edu

ABSTRACT

This research is motivated by the dominance of the Google search engine and the bias that it may introduce to the users' perception of the Web. According to a recent study, 75% of keyword searches on the Web are handled by Google [17]. Given that Google returns currently "popular" pages at the top of search results, are we unfairly penalizing newly created pages that are not yet very well known? Is there a better way of measuring the "quality" of a page than using the "popularity" of the page? In this paper, we propose a new definition of page quality and develop a practical way of measuring the proposed page quality based on the evolution of the Web link structure. We prove that our proposed quality estimator measures the quality of a page well through a theoretical analysis of a reasonable Web user model. We also present our experimental results that show the potential of our estimator in measuring the page quality. 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

Since its founding in 1998, Google has become the dominant search engine on the Web. According to a recent estimate [17], about 75% of Web searches are being handled by Google directly or indirectly. For example, in addition to the keyword queries that Google gets directly from its sites, all keyword searches on AOL are routed to Google. It is this startling dominance that led one Internet commentator to conclude that, essentially, "if your page is not indexed by Google, your page does not exist on the Web [16]." While this statement may be an exaggeration, it contains an alarming bit of truth. To find a page on the Web, many users go to Google (or another search engine that uses Google results), 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. A page not indexed by Google or ranked poorly by Google is therefore not likely to be viewed by many Web users.

Our research is motivated by this dominance of Google and the bias that it may introduce. Is the people's perception of the Web influenced by Google? What kind of and how much bias does it introduce? Is there a way to reduce this bias? Our research is particularly concerned about Google's ranking of Web pages and the bias induced from this ranking.

While Google takes more than 100 factors into account in determining the final ranking of a page [9], the core of their ranking algorithm is based on a metric called PageRank [18, 5]. PageRank is essentially a "link-popularity" metric, where a page is considered more important if the page is linked to by many other pages on the Web.[1] Roughly speaking, Google puts a page at the top in 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. The effectiveness of Google's search results and the adoption of PageRank by major search engines [23] strongly indicate that PageRank is an effective ranking metric for Web searches.

It is important to understand the distinction between the *importance* or *quality* of a page and the *relevance* of a particular Web document to a particular search. The relevance is a quantity that depends heavily on the particular search issued by a user. In contrast, the importance or quality of the document is a quantity that can be computed at crawl time and could be seen as intrinsic to the document itself. It is this intrinsic document quality with which this paper is concerned.

The core assumption of PageRank is that pages that are very popular are the pages of highest quality. But one important problem is that PageRank is based on the *current* popularity of a page. Since currently-popular

[1] A more precise description of the PageRank metric is provided in Section 3.

pages are repeatedly returned by search engines as the top results, they are also the easiest for users to discover, which increases their popularity further. In contrast, a currently-unpopular page is often not returned by search engines, so few new links will be created to the page, keeping page's ranking down. This "rich-get-richer" phenomenon can be particularly problematic for the high-quality yet currently-unpopular pages. 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. It is clearly unfortunate (both for the author of the new page and the overall Web users) that important and useful information is being ignored simply because it is new and has not had a chance to be noticed. It is here we see this core assumption of PageRank violated.

Now that we have identified this flaw in the PageRank metric, can we avoid this problem? That is, is there a way to rank pages based on their quality and not simply on their popularity?

At the core of this problem lies the question of page quality. What do we mean by the quality of a page? Without a good definition of page quality, it is difficult to measure how much bias PageRank induces in its ranking and how well other ranking algorithms capture the quality of pages.

In this paper we first try to clarify the notion of page quality and introduce a formal definition of *page quality*. Our quality metric is based on the idea that if the quality of a page is high, when a Web user reads the page, the user will probably like the page (and create a link to it). So we define the quality of a page as the probability that a Web user will like the page enough to create a link to it when he reads the page. We then propose a *quality estimator*, or a practical way of estimating the quality of a page. Our quality estimator analyzes the changes in the Web link structure and uses this information to estimate page quality. We theoretically show that our estimator can measure the exact quality of pages based on a simple and reasonable Web model. We also present our experimental results that show the potential of our estimator in measuring the quality of a page. In 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)
- We show that Google's PageRank measures our formal definition of page quality very well in certain conditions. We also argue that Google's PageRank is heavily 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. We provide the intuition and the derivation of our proposed estimator. (Sections 5 through 7)
- We conduct an experiments 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

[22] 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 [27], the vector-space model [21] and the probabilistic model [20, 7] are some of the well known models developed in this context. Some of these models (particularly the vector-space model) were adopted by many of the early Web search engines. This work is, however, geared towards measuring the relevance of a page rather than its quality.

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 [13] and PageRank [18] 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, 11, 12] 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. [10] studies personalization of the PageRank metric by giving different weights to pages. [25] proposes a modification of PageRank equation to tailor it for Web administrators. [23] 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 [8] and measuring the object proximity.

There exists a large body of work that investigates the properties of the Web link structure [3, 4, 6, 19]. For example, [6] shows that the global link structure of the Web is similar to a "bow tie." [3, 6] shows that the number of in-bound or out-bound links follow a power-law distribution. [4, 19] propose potential models on the Web link structure.

The probabilistic model [7, 20] developed in the IR community is similar to our quality metric in that both definitions take a probabilistic approach. The probabilistic model, however, measures that 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). Many of the important 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 $p_1, \ldots, p_m$. Let c_j be the number of links going out of page p_j.[2] Then, the PageRank of page p_i is given by

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

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.

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.[3] 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 [23], 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 [15].

We should note that the PageRank algorithm described here is unlikely to be exactly the technique used by Google today. Since the founding of Google as a private company, the development of PageRank has continued, and likely now contains corrections or other measures to deal with "search engine optimizers" who interfere with PageRank. While for the purposes of this paper we will use the basic PageRank algorithm described in this section, we could easily include more advanced techniques.

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.

While quite effective, one significant flaw of PageRank is that it is inherently biased against unpopular pages. 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, there exist only a few (or no) links to the page and thus search engines never return the page or give it very low rank. Because search engines do not return it, few users ever see this page, so the popularity of the page does not increase very much. It may take a very long time for this new page to be discovered by enough users to be ranked highly by the search engine.

To avoid this problem, is there a way to measure the quality of a page and somehow promote the high-quality (yet not very popular) pages? The first challenge to the problem is the notion of "page quality." What do we mean by page quality? How can we quantify it?

We note that page quality can be a very subjective notion; different people may have completely different quality judgment on the same page. One person may regard a page very highly while another person may consider the page completely useless. Given this subjectivity, is it possible to come up with a reasonable definition of page quality that on which most people can agree?

In this paper, we propose to quantify the *quality* of a page as the probability that a random Web user will like the page enough to create a link to it once that user discovers the page for the first time.

[2] If a page has no outgoing link, we assume that it has outgoing links to every single Web page.

[3] When the user reaches a page with no outlinks, he jumps to a random 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 and create a link to the page p given that the user discovers the page for the first time. Mathematically,

$$Q(p) = P(L_p|A_p)$$

where A_p represents the event that the user becomes newly aware of the page p and and L_p represents the event that the user likes the page and creates a link to p. □

Given this definition, we can hypothetically measure the quality of page p by showing p to *all* Web users (or to a sample of Web users) and getting the users' feedback on whether they like p or not (or by counting how many people create a link to p). 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 believe that this is a reasonable way of defining page quality given the subjectivity of page quality. When individual users have different opinions on the quality of a page, it is reasonable to consider a page to be of higher quality if more people are likely to "vote for" the page.

Under this 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 to be of higher quality under our definition simply because the movie "Star Wars" is more popular than "Latino." This again is the central issue of relevance versus quality. Any search engine must also employ techniques to narrow a search to a particular set of relevant documents. It is only once this set of documents (say, pages on the movie Latino) has been identified that the quality or importance metric is used to rank the pages, so 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 this "topic bias" does not hurt the effectiveness of a search engine using our metric.

Note that the current popularity (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, the number of people who like the page (and create a link to it) 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. In the next section, we discuss how we can measure the quality of a page using the evolution of the Web link structure.

5. QUALITY ESTIMATOR: INTUITION

Although we arrived at the quality estimator theoretically based on a user visitation model, before we delve into this detailed analysis we will in this section provide a more intuitive explanation. This way, the more technical analysis in Sections 6 and 7 will be easier to understand.

How can we measure the quality of a page without asking for user feedback? Given that the quality of a page is the fraction of the Web users who create a link to the page once the users visit the page, we may suspect that the time derivative of the link count (or PageRank) of the page may provide some information of its quality. The difficult question is exactly how we should use the time derivative to correctly estimate the quality.

Our main idea for quality measurement is as follows: The quality of a page is how many users will like a page and create a link to the page when they discover it. Therefore, if two pages are discovered by the same number of people during the same period, more people will create a link to the higher-quality page. In particular, the increase in the link count (or popularity) is directly proportional to the quality of a page. Thus, by measuring the increase in popularity, not the current popularity, we may estimate the page quality more accurately. We can use here any measure of popularity. We will use PageRank for the purposes of this paper because of its success as a popularity metric, but we could just as easily substitute the number of links.

There exist two problems with this approach. The first problem is that pages are *not* visited by the same number of people. A popular page will be visited by more people than an unpopular page. Even if the quality of pages p_1 and p_2 are the same, if page p_1 is visited by twice as many people as p_2, it will get twice as many new links as p_2. To accommodate this fact, we need to divide the popularity increase by the number of visitors to this page. Given that current PageRank (or popularity) captures the probability that a random Web surfer arrives at a page, we may assume that the number of visitors to a page is proportional to its current popularity. We thus divide the increase in the popularity by the current PageRank to measure quality.

The second problem is that the popularity of an already well-known page may not increase very much because it is already known to most Web users. Even though many users visit the page, they do not create any more links to the page because they already know about it and have already created links to it. Therefore, if we estimate the quality of a well-known page simply based on the increase in the popularity, the estimate may be lower than its true quality value. We avoid this problem by considering both the current popularity of the page and the increase in the popularity. More precisely, we measure the quality of page through the following formula:

$$Q(p) \approx C \cdot \frac{\Delta PR(p)}{PR(p)} + PR(p) \qquad (1)$$

Here, the first term, $\frac{\Delta PR(p)}{PR(p)}$, estimates the quality of a

page by measuring the relative increase in its PageRank.[4] The second term, $PR(p)$, is to account for the well-known pages whose PageRank do not increase much because they are already so well known. When the PageRank of a page has saturated, we believe that the saturated PageRank value reflects the quality of the page: a higher-quality page is eventually linked to by more pages. The constant C in the formula decides the relative weight that we give to the increase in PageRank and to the current PageRank.

Note that we can measure the values in the above formula in practice by taking multiple snapshots of the Web. That is, we download the Web multiple times, say twice, at different times. We then compute the PageRank of every page in each snapshot and take the PageRank difference between the snapshots. Using this difference and the current PageRank of a page, we can compute its quality value.

In the next two sections, we present our original theoretical analysis that led us to the above estimator.[5] In Section 6 we explain our model on how Web users visit pages. In Section 7, we analyze how the popularity of a page evolves over time under this model and use the result to obtain the quality estimator.

6. OUR USER-VISITATION MODEL

6.1 Basic definitions

Before we explain our user-visitation model, we introduce three definitions that are important to understand our model. First, we introduce two notions 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, say, 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.

We note that the exact popularity of a page is difficult to measure in practice. However, we may substitute any popularity metric, such as PageRank, as a surrogate to its 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 the Web users who is 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. 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.*

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

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. In our model, we assume that the quality $Q(p)$ is an inherent property of p 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.

For 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.

6.2 User-visitation model: two hypotheses

We now explain two core hypotheses of our user-visitation model. 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 surfs the Web. Given this interpretation, it

[4] We may replace $PR(p)$ in the formula with the number of links

[5] In fact, the intuition and the potential problems that we described in this section was gained with a hindsight after we arrived at the final form of the estimator.

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

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 common to all pages. □

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.* □

7. THEORETICAL DERIVATION OF QUALITY ESTIMATOR

The goal of this section is to investigate the user-visitation model described in the previous section and see how we may measure the quality of page p by observing the evolution of its popularity. For this purpose, we first analyze the popularity evolution of a page under the model.

7.1 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, out of these new users, $Q(p)$ fraction 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,

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

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. Combined together, the lemmas show that we can compute the current popularity of a page from its past popularity. The following theorem shows the popularity evolution given its initial popularity.

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. □

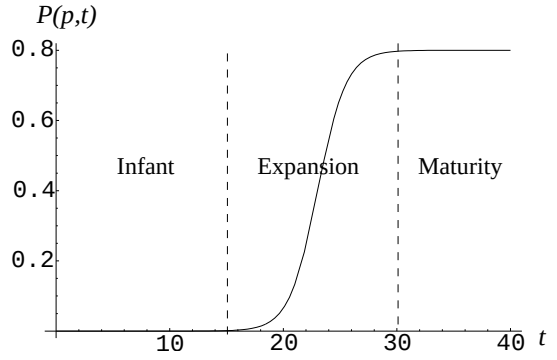

Figure 1: Time evolution of page popularity

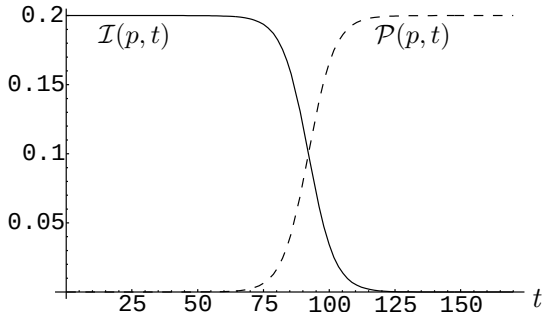

Figure 2: Time evolution of $\mathcal{I}(p,t)$ and $\mathcal{P}(p,t)$ as predicted by the model.

In the rest of this paper, we skip the proofs due to their length and complexity. Proofs are not important to understand the core idea of the paper, but interested readers may read Section 11 for the proofs.

Based on the result of the above theorem, we show an example of the popularity evolution of a page in Figure 1. We assumed $Q(p) = 0.8$, $n = 10^8$, $r = 10^8$ 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 [14]).

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 $t \to \infty$, $\mathcal{P}(p,t) \to Q(p)$.* □

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 when the page has just been created. 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. In the next subsection, we check whether the time derivative of the popularity is a better estimator of the page quality.

7.2 Derivation of quality

Our main idea for better quality estimation is to use the popularity increase of the page as the quality estimator. To check the validity of this idea, we take the time derivative of $\mathcal{P}(p,t)$ and get the following lemma.

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, $1 - \mathcal{A}(p,t)$.*

$$Q(p) = \left(\frac{n}{r}\right) \frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)\,(1 - \mathcal{A}(p,t))} \qquad (2)$$

□

In the above equation, note that two main factors, $d\mathcal{P}(p,t)/dt$ and $\mathcal{P}(p,t)$, are measurable in practice while $1 - \mathcal{A}(p,t)$ is not. That is, we can measure $d\mathcal{P}(p,t)/dt$ by downloading the Web multiple times and measuring popularity increase of p. $\mathcal{P}(p,t)$ can also be measured from its current popularity. $\mathcal{A}(p,t)$ is, however, difficult to measure because we do not know how many users are aware of p unless we know when p was first created and how many users have visited it so far. 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. For convenience, we refer to $\left(\frac{n}{r}\right) \frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)}$ as the *relative popularity increase*, $\mathcal{I}(p,t)$.

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. We obtained this graph analytically using the equation of Theorem 1. 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.

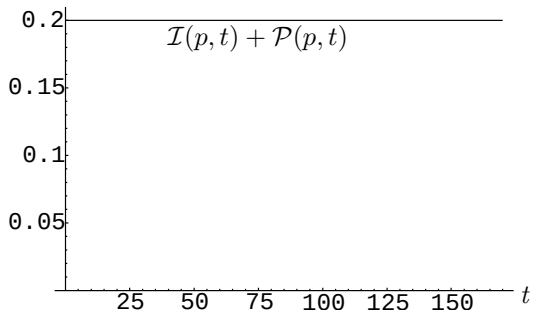

Figure 3: Time evolution of $\mathcal{I}(p,t) + \mathcal{P}(p,t)$.

From the graph, we can see that the relative popularity increase $\mathcal{I}(p,t)$ measures the quality of the page $Q(p)$ very well in the beginning when the page was just created ($t < 70$). During this time, $\mathcal{I}(p,t) \approx 0.2 = Q(p)$. As time goes on, however, the relative popularity increase $\mathcal{I}(p,t)$ loses its merit as the estimator of $Q(p)$. $\mathcal{I}(p,t)$ gets much smaller than $Q(p)$ for $t > 120$. This result is reasonable because when most users on the Web are aware of the page, the popularity of the page cannot increase any further, so the popularity-increase-based quality estimator will be much smaller than $Q(p)$. In contrast, the popularity $\mathcal{P}(p,t)$ works very poorly as the estimator of $Q(p)$ in the early stage of a page ($t < 70$), but is a good estimator of $Q(p)$ when t is large ($t > 120$).

From the above discussion, we can see that $\mathcal{I}(p,t)$ and $\mathcal{P}(p,t)$ are complementary to each other as the quality estimator of a page. When $\mathcal{P}(p,t)$ does not work well as the quality estimator, $\mathcal{I}(p,t)$ does. When $\mathcal{I}(p,t)$ does not, $\mathcal{P}(p,t)$ does. In fact, from the shape of the two curves 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 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 $\mathcal{P}(p,t)$.*

$$Q(p) = \mathcal{I}(p,t) + \mathcal{P}(p,t) \qquad \square$$

Based on the result of Theorem 2, we define $\mathcal{I}(p,t) + \mathcal{P}(p,t)$ as the *quality estimator* of p, $\mathcal{Q}(p,t)$:

$$\begin{aligned} \mathcal{Q}(p,t) &= \mathcal{I}(p,t) + \mathcal{P}(p,t) \\ &= \left(\frac{n}{r}\right)\left(\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)}\right) + \mathcal{P}(p,t) \end{aligned} \qquad (3)$$

Intuitively, the above equation shows that under our user-visitation model we can exactly estimate the quality of a page by measuring its relative popularity increase and current popularity, which in turn can be measured by downloading the Web multiple times.

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 new 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 for 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, requiring a careful selection of a representative user group and Web pages and a rigorous way of ensuring the unbiasedness of the collected user feedback. Recognizing this challenge, the IR community has collaboratively constructed a standard evaluation corpus, called TREC [24], which also includes a special sub-collection 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.

Given this difficulty, we take an alternative approach to evaluating the potential of our quality estimator. Our main idea for evaluation is that the popularity or PageRank of a page is a reasonably good estimator of its quality if the page has existed on the Web for a long period (Corollary 1). Thus, if we can wait long enough, the future PageRank of a page will be close to its true quality. This means that if our quality estimator estimates the quality of the page well, the estimated page quality from today's Web should be closer to the future PageRank than the current PageRank. In other words, our quality estimator should be a better "predictor" of the future PageRank than the current PageRank. Based on this idea, we capture multiple snapshots of the Web, compute page quality, and compare today's quality value with the PageRank values 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 (which can be measured reliably only through a large-scale user study), we believe that this comparison, at the very least, will show the potential of our estimator. In the remainder of this section, we describe our experimental setup and the results obtained from our experiment.

8.1 Experimental Setup

Due to our limited network and storage resources, we had to restrict our experiments to a relatively small subset of the Web. 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 4. Roughly, the first three snapshots were taken with one-month interval between them and the last snapshot was taken four months after the third snapshot. We refer to the time of each snapshot as t_1, t_2, t_3 and t_4. The first three snapshots were used to compute the quality of pages and the last snapshot was used as the "future" PageRank.

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. For the computation, we used 1 as the initial PageRank value of each page.

8.2 Quality and future PageRank

Using the collected data, we estimated the quality of a page based on the PageRank increase between t_1 and t_3. We then compared the estimated quality to the PageRank at t_4 and measured the difference. In estimating page quality, we first identified the set of pages whose PageRank values had consistently increased (or decreased) over the first three snapshots (i.e., the pages with $PR(p,t_1) < PR(p,t_2) < PR(p,t_3)$). For these pages, we computed the quality through the following formula:

$$Q(p) = 0.1 \cdot \left[\frac{PR(p,t_3) - PR(p,t_1)}{PR(p,t_1)} \right] + PR(p,t_3)$$

That is, we computed the PageRank increase by taking the difference between t_1 and t_3 ($\Delta PR(p) = PR(p,t_3) - PR(p,t_1)$) and dividing it by $PR(p,t_1)$. We then added this number to $PR(p,t_3)$ to estimate the page quality. As the constant factor C in Equation 1, we used the value 0.1.[6] Note that our quality estimator becomes the same as the current PageRank if the PageRank of a page does not change between t_1 and t_3. Since the majority of pages did not show a significant change in PageRank values, in the remainder of this section, we report our results only for the pages whose PageRank values changed more than 5% between t_1 and t_3. By limiting to these pages, we can see the potential of our quality estimator more clearly.

[6]The value 0.1 showed the best result out of all values that we tested. Small variations in the constant did not affect our result significantly.

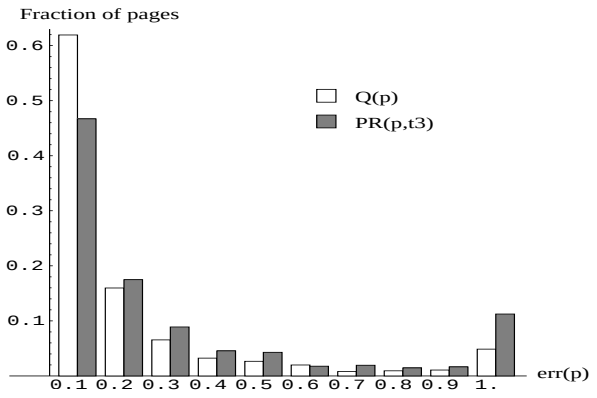

Figure 5: Histogram of relative errors

In order to quantify how well $Q(p)$ predicts the "future" PageRank $PR(p,t_4)$ compared to the "current" PageRank $PR(p,t_3)$, we compute the average relative "error" between $Q(p)$ and $PR(p,t_4)$ and between $PR(p,t_3)$ and $PR(p,t_4)$. That is, we compute the relative error

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

for the pages and compare their average errors.

From this comparison, we could observe that the average relative error is significantly smaller for $Q(p)$ than $PR(p,t_3)$. The average error was 0.32 for $Q(p)$ while it was 0.78 for $PR(p,t_3)$. That is, our quality estimator $Q(p)$ predicted the future PageRank *twice* as accurately as $PR(p,t_3)$ on average. Assuming that the PageRank at t_4 is closer to the quality of pages, this result strongly indicates that our estimator measures the quality much more accurately than the current PageRank.

In Figure 5, we report more detailed result from this comparison. In the graph, we show the distribution of the relative errors for $Q(p)$ and $PR(p,t_3)$. We counted how many pages had the relative error between 0 and 0.1, 0.1 and 0.2, etc., and plotted the histogram. The white bars correspond to the histogram of $Q(p)$ and the grey bars correspond to $PR(p,t_3)$. The bars labeled as 0.1 correspond to the error range between 0 and 0.1, the bars labeled as 0.2 correspond to the range between 0.1 and 0.2, etc. When the error was larger than 1, we put them into the last bin labeled as 1. The vertical axis shows the fraction of pages within the given error range. From the histogram, we can see that our quality estimator $Q(p)$ shows significantly smaller error than $PR(p,t_3)$. For example, from the first bars of the graph we can see that $Q(p)$ showed less than 0.1 relative error for about 62% of the pages, while $PR(p,t_3)$ showed similar error only for 46% of the pages. Also,

Figure 4: The timeline that our four snapshots were taken

from the last bars of the graph, we can see that $Q(p)$ showed relative error larger than 1 only for 5% of the pages, while $PR(p, t_3)$ showed similar error for over 10% of the pages.

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 proposed a practical way of estimating the quality of a page using the evolution of the Web link structure. We then theoretically justified our quality estimator and experimentally showed the potential of our quality estimator.

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 [13, 18] 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.

- *Decreasing popularity*: Our user-visitation model predicts that the popularity of a page only increases over time (Figure 1). However, many pages in our dataset showed consistent decrease in their PageRanks. In order to explain these pages, a revision in our user-visitation model is necessary. We expect that we may explain popularity decrease by modeling the fact that some users may "forget" some of the pages that they visited.

- *Fluctuation in PageRank*: Even further, during the analysis of our experimental result, we observed fluctuation of PageRanks for many of the pages that we downloaded. For example, the PageRank values for a number of pages went up from t_1 to t_2 and went down again from t_2 to t_3. For these pages, we assumed that $\mathcal{I}(p, t) = 0$ for our quality estimator because when their PageRank values oscillate, it is difficult to estimate this part. Again, our model does not handle these pages.

- *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 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. REFERENCES

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

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

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

[4] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the International World-Wide Web Conference*, April 1998.

[6] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web: experiments and models. In *Proceedings of the International World-Wide Web Conference*, May 2000.

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

[8] Roy Goldman, Narayanan Shivakumar, Suresh Venkatasubramanian, and Hector Garcia-Molina. Proximity search in databases. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 26–37, 1998.

[9] Google information for webmasters. Available at `http://www.google.com/webmasters/`.

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

[11] Sepandar Kamvar, Taher Haveliwala, and Gene Golub. Adaptive methods for the computation of pagerank. In *Proceedings of International Conference on the Numerical Solution of Markov Chains*, September 2003.

[12] Sepandar Kamvar, Taher Haveliwala, Christopher Manning, and Gene Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the International World-Wide Web Conference*, May 2003.

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

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

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

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

[17] Search engine market research by onestat.com. Brief summary is available at `http://www.onestat.com/html/aboutus_pressbox21.html`, May 2002.

[18] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry 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`.

[19] David M. Pennock, Gary W. Flake, Steve Lawrence, Eric J. Glover, and C. Lee 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.

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

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

[22] Gerard Salton and Michael J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, 1983.

[23] John A. Tomlin. A new paradigm for ranking pages on the world wide web. In *Proceedings of the International World-Wide Web Conference*, May 2003.

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

[25] Ah Chung Tsoi, Gianni Morini, Franco Scarselli, Markus Hagenbuchner, and Marco Maggini. Adaptive ranking of web pages. In *Proceedings of the International World-Wide Web Conference*, May 2003.

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

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

11. PROOFS

Proof for Theorem 1 From Lemmas 1 and 2,

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

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 [26]. 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

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

Proof for Corollary 1 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,

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

Proof for Lemma 3 By differentiating the equation in Lemma 1, we get

$$\frac{d\mathcal{P}}{dt} = \frac{d\mathcal{A}}{dt}Q(p). \tag{9}$$

From Lemma 2,

$$\begin{aligned}\frac{d\mathcal{A}}{dt} &= -\frac{d}{dt}e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt} \\ &= -\left(e^{-\frac{r}{n}\int_0^t \mathcal{P}(p,t)dt}\right)\left(-\frac{r}{n}\mathcal{P}(p,t)\right) \\ &= (1 - \mathcal{A}(p,t))\left(\frac{r}{n}\mathcal{P}(p,t)\right).\end{aligned} \tag{10}$$

From Equations 9 and 10, we get

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

Proof for Theorem 2
If we multiply Equation 2 by $1 - \mathcal{A}(p,t)$, we get

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

The right-hand side of the above equation is $\mathcal{I}(p,t)$. The left-hand side is

$$Q(p) - Q(p)\cdot\mathcal{A}(p,t) = Q(p) - \mathcal{P}(p,t).$$

Therefore,

$$Q(p) - \mathcal{P}(p,t) = \mathcal{I}(p,t).$$ ■