

# Online Reputation Systems: The Cost of Attack of PageRank

Andrew Clausen

December 4, 2003

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Reputation Systems</b>	<b>4</b>
2.1	Definitions . . . . .	4
2.1.1	Defining Trust . . . . .	4
2.1.2	Defining Reputation . . . . .	4
2.1.3	Defining Online Reputation Systems . . . . .	5
2.1.4	Reputation Systems and Reputation Metrics . . . . .	5
2.2	Brief Survey of Reputation Systems . . . . .	5
2.3	Sybil Attack . . . . .	6
2.4	The Need for Complaints . . . . .	7
2.5	Problems with Complaints . . . . .	7
2.6	Scope of Reputation . . . . .	7
2.7	Conclusion . . . . .	8
<b>3</b>	<b>Defining PageRank</b>	<b>9</b>
3.1	Markov Theory . . . . .	9
3.1.1	Markov processes . . . . .	9
3.1.2	Convergence of Markov processes . . . . .	10
3.1.3	Markov Process Isomorphisms . . . . .	11
3.1.4	Lumpable Markov processes . . . . .	12
3.1.5	Summary of Markov Theory . . . . .	13
3.2	Random Surfer Model . . . . .	13
3.2.1	Intuition . . . . .	13
3.2.2	Formal definition . . . . .	13
3.2.3	Algebraic PageRank . . . . .	16
3.3	Definition Confusion and Contradiction . . . . .	17
3.3.1	<i>PageRank1Wrong</i> as defined in [Brin and Page, 1998] . . . . .	17
3.3.2	<i>PageRank2</i> as defined in [Page et al., 1998] . . . . .	19
3.3.3	Summary . . . . .	20
<b>4</b>	<b>PageRank Cost of Attack</b>	<b>21</b>
4.1	PageRank Cost of Attack Theorem . . . . .	21
4.1.1	Overview . . . . .	22
4.1.2	Mathematical Proof . . . . .	24
4.2	Cost of Attack of Google . . . . .	31
4.2.1	Adapting the Theorem . . . . .	31
4.2.2	Estimating the Cost of Attack . . . . .	32
4.2.3	Comparison with Empirical Cost of Attack . . . . .	33
4.3	Limitations . . . . .	34

4.4	Implications . . . . .	35
4.5	Future Work . . . . .	35
<b>5</b>	<b>Contribution</b>	<b>36</b>
<b>6</b>	<b>Conclusions</b>	<b>37</b>

# Chapter 1

## Introduction

Trust is fundamental to many activities involving several people. Consider the case of an employer examining a potential employee's university grade average. Perhaps hundreds of individuals contributed to the decision behind this number. Why would any employer trust such a huge collaborative decision about an individual?

The employer might have hired productive employees from the same university. Additionally, the employer probably has confidence in the accountability structure within the university. Academics, tutors and other university employees would be subject to harsh punishments if they offered higher grades for 'private tuition'. Or, perhaps the employer has read articles published by the university's academics.

eBay is a web site that allows individuals to buy and sell goods through online auctions. The highest bidder sends the payment, trusting that the vendor will send the goods as promised. Without the feedback system, this trust would probably be misplaced. Unlike the employer seeking an employee, the bidder can not verify the location of the vendor, how long the vendor has been trading and probably does not know anyone who has traded with the vendor. Moreover, if the vendor lived in a different state or country, legal recourse would not likely be a viable option.

eBay's feedback system allows traders to leave positive and negative recommendations after transactions. It aggregates this mountain of information into a single reputation score for every trader. This reputation score is the difference between boom and bust. Without its reputation system, eBay would be an offer you couldn't risk. Instead, \$US14.87 billion of goods were traded on eBay in 2002.<sup>1</sup>

Despite the success of eBay's reputation system, anecdotes of fraud and deception continue[Calkins, 2001]. Perhaps this fraud could be reduced by installing a better reputation system. Moreover, a better reputation system might make it possible for traders to establish trust where they could not reasonably do so before.

Perhaps the best way to evaluate a reputation system is to measure the cost of attacking it. That is, measuring the cost of increasing an individual's reputation through deception. If this cost is high, then traders can be confident that reputation scores were obtained legitimately. Moreover, they can assess whether a trader's reputation is sufficiently high to be trusted. Clearly, a deceptive trader may well pay to attack a reputation system if the payoff from a trade were large enough.

Therefore, an important goal of reputation systems research is to design a system that has a high cost of attack. I argue that Google's PageRank [Page et al., 1998] algorithm is a promising foundation for reputation systems, since all information is weighted by the reputation of the source. Therefore, the aim of this thesis is to determine the cost of attack of PageRank.

In *Chapter 2*, I review the existing literature in Reputation Systems. In *Chapter 3*, I review the PageRank literature. I also review the Markov theory literature that provides the mathematical foundations for PageRank. In *Chapter 4*, I develop a mathematical theory to measure the cost of attack of reputation system that use PageRank. I then apply this theory to measure the cost of attack of Google, and compare this to empirical evidence. Finally, I summarize my contributions in *Chapter 5*, and conclude with *Chapter 6*.

---

<sup>1</sup><http://pages.ebay.com/community/aboutebay/overview/index.html>

## Chapter 2

# Reputation Systems

In this chapter, I review the reputation systems literature. I first give some working definitions of terminology. Then, I review existing reputation systems and their susceptibility to attack. I argue that PageRank has the desirable property of weighting recommendations by reputation, and that eBay has the desirable capability of registering complaints. I conclude that it would be desirable to combine these two properties.

### 2.1 Definitions

*Trust* and *reputation* are difficult to define because they are subtle concepts with many complexities. As a result, most of the Reputation Systems literature avoid this issue. Nonetheless, a poor working definition is better than no definition.

In this section, I choose definitions for *trust*, *reputation* and *online reputation system* from the literature. I also propose a new term: *reputation metric*.

#### 2.1.1 Defining Trust

[Marsh, 1994] surveys definitions of *trust* from the social sciences, and argues that trust is not sufficiently well understood to be defined. As a result, most literature on reputation systems including [Zacharia et al., 1999, Resnick et al., 2000, Calkins, 2001, Kamvar et al., 2003] omit any definition.

However, [Mui et al., 2001] define trust as ‘a subjective expectation an agent has about another’s future behaviour based on the history of their (*the other’s*) encounters’.

I will also use this definition as it is adequate for the purposes of this thesis.

#### 2.1.2 Defining Reputation

[Resnick et al., 2000] consider reputation to be the community opinion of a subject. They argue that reputation is useful to an individual deciding how much to trust for two important reasons. Firstly, because previous interactions indicates what future behaviour might be like. Secondly, if an individual’s recommendation or complaint is likely to be taken seriously by a community, then the fear of a bad reputation can introduce accountability. Moreover, accountability can lead to trust.

I think this is a clean separation of trust and reputation. These definitions matches intuition from everyday experience in the world. It allows scope for some members of a community’s opinions to carry more weight than others. It also does not require that all gossip or feedback be correct, but possibly constructed strategically to deceive.

However, the definition of reputation has not been widely agreed. [Mui et al., 2001] defines reputation as ‘a social quantity based on actions by a given agent  $a_i$  and observations made by others in an “embedded social network” that  $a_i$  resides in’. The same authors define it in [Mui et al., 2002] as “a perception that an

agent has of another’s intentions and norms”. The second definition seems to be at odds with the first one, more in line with trust than reputation. The first definition seems to imply that at all agents have direct access to the observations of all other agents.

I use the term *reputation* to refer to group or community opinion.

### 2.1.3 Defining Online Reputation Systems

Currently, the literature refers to *reputation systems* rather than *online reputation systems*. While *reputation system* is a reasonable abbreviation, it can cause confusion for researchers outside of computing. I use these terms interchangeably in this thesis.

[Resnick et al., 2000] defines a *reputation system* as “a system that collects, distributes and aggregates feedback about participants’ behaviour”.<sup>1</sup> This describes how a reputation system works, but not what we expect from them. Therefore, I suggest *reputation system* be defined as ‘a system that assigns reputations to participants’. Note that I intend reputation to mean ‘group opinion’, as I defined it in the previous section.

In particular, this definition probably excludes collaborative filtering systems. A collaborative filtering system such as GroupLens [Resnick et al., 1994] allows large communities of users to rate items such as movies, music and newsgroup postings. Once a user has registered their ratings for a few items, their taste can be matched against other users’ preferences, allowing predictions about ratings for items a user has never encountered before.

While collaborative filtering systems do compute reputation scores of items, they assume that all participants are trustworthy, albeit with different tastes. In particular, they do not consider the reputation of the recommenders. Some authors including [Zacharia et al., 1999] have described collaborative filtering systems as reputation systems. However, I will argue that ignoring the reputation of participants leads to skew results that do not match community opinion. It could be argued that a collaborative filtering system finds the sub-community that a user fits into, and then evaluates the opinion of that sub-community. This selection of appropriate sub-community is not based on reputation, so I argue collaborative filtering systems can only be considered one component of a reputation system.

### 2.1.4 Reputation Systems and Reputation Metrics

The term *Reputation systems* has been used by [Mui et al., 2002], [Sabater and Sierra, 2002] and [Resnick et al., 2000] to refer to complete systems such as eBay and Google rather than the algorithms underpinning them.

Two terms have been suggested for an algorithm or mathematical function that computes reputation scores: *reputation model* in [Mui et al., 2002] and [Sabater and Sierra, 2002]; and *trust metric* in [Levien, 2003].

I think both terms are misleading. *Model* implies that we are trying to measure and understand how real people and communities behave. However, I think these metrics are normative, not descriptive, because we are trying to develop social norms that will lead to greater accountability. Secondly, I think *trust metric* is misleading since these algorithms (including [Levien, 2003]) measure reputation, not trust. Therefore I will take the liberty of calling these algorithms *reputation metrics*.

## 2.2 Brief Survey of Reputation Systems

In this section, I briefly survey reputation systems.

**eBay** is an online auction market place that allows users to give feedback about each other. After each transaction, the parties may rate each other as good, bad or neutral. eBay then reports a reputation score that is the sum of all positive feedback minus the sum of all negative feedback.

This is simple. It does not take the parties’ prior reputation into account. For example, a recommendation from a highly reputable computer supplier selling thousands of machines every year is not considered any more significant than a recommendation from a first-time user of eBay. Moreover, it does not consider recommendations from users reputed to have relevant expertise any differently.

---

<sup>1</sup>The authors were possibly just describing rather than attempting to define *reputation system*.

**Google** [Brin and Page, 1998] is a popular search engine that uses the **PageRank** reputation metric to evaluate the reputation of web pages.

PageRank weights recommendations by the source's reputation score. PageRank requires users to specify which pages they trust. These trusted pages are then granted the reputation votes that then propagate through the network.

Google uses PageRank scores to decide which pages should appear prominently in search results. It is widely believed [Levien, 2003] that Google allocates the initial votes to either top-level web pages registered with the Domain Name System, or possibly those registered with the Open Directory project.<sup>2</sup>

PageRank has been described as "Google's original sin" by [www.google-watch.org](http://www.google-watch.org). I don't address their arguments in this thesis.

**EigenTrust** [Kamvar et al., 2003] is a research reputation system for peer to peer networks that is based on PageRank. EigenTrust aims to reward peers that behave well in a peer to peer network. It does not attempt to collect negative reputation information. It allocates the initial votes to a set of *a priori* trusted peers that all peers trust.

**Advogato** [Levien, 2003] is a free software community website with a rating system for free software developers. Developers can certify each other as Apprentice, Journeyer or Master. The reputation metric is based on network flows where each developer has two nodes in a network, and each certificate is an edge between two developers' nodes. In fact, there are three networks - one for each level. Master certificates appear in all three networks, and Journeyer only in Journeyer and Apprentice networks, and so on. The reputation of a developer is the best network for which the developer's nodes have a non-zero flow. Unfortunately, maximum flows of networks are not unique. Therefore allocation of reputation is partially arbitrary.

**Regret** [Sabater and Sierra, 2002] is a research reputation system that estimates the reputation (community opinion) of an individual by taking a member of the community's opinion. Regret attempts to select the most appropriate witness's recommendation according to relevance of the witness's interaction, conflict of interest and social structure. It does not apply this witness selection procedure recursively - that is witnesses would not necessarily have high reputation scores. Even though this is only strictly taking an individual's opinion, it is claimed that this is a good approximation of the opinion of a community.

[Mui et al., 2001] proposes a model of trust and reputation that I will call **Mui**. It gives a statistical function for computing reputation. The model computes reputation from the history of actions (cooperate and defect) between all people rather than recommendations or complaints. It provides no method for critical evaluation of history information; it is assumed that all history information is reliable.

[Aberer and Despotovic, 2001] proposes a peer to peer reputation system that I will call **Aberrer**. Unlike the other reputation systems surveyed here, Aberrer only supports complaints. Peers can only acquire negative reputations. Both issuing and receiving complaints worsens a peer's reputation.

## 2.3 Sybil Attack

Sybil attack [Douceur, 2002] is named after the famous (alleged) multiple personality case of Sybil Dorsett. A Sybil attack is an attack on a system that involves a single person voting many times under the guise of multiple identities, such as branch-stacking in political preselection elections.

It has been conjectured by [Page et al., 1998] and [Levien, 2003] that **PageRank**'s initial vote system and recommendation weighting work together to resist Sybil attack. They claim that identities created en masse can only inherit their reputations from those recommending them. This claim extends to both **Google** and **EigenTrust**.

However, [Bianchini et al., 2004] proved that PageRank is vulnerable to Sybil attack if the initial votes are allocated uniformly. On the world wide web, Sybil attacks by creating lots of web pages are commonly called 'link farms'.

**Advogato** only permits reputation to flow through paths from a seed. Only upstream recommendation structure has any impact, so it resists Sybil attack. Like PageRank, Advogato limits the impact any person

---

<sup>2</sup>Open Directory Project: <http://www.dmoz.org>

can have by the amount of reputation flow the person has. [Levien, 2003] describes this as the *bottleneck property*.

**eBay** only permits feedback to be published after transactions. Naturally, eBay takes a small percentage of each transaction, so each piece of feedback involves some money being paid to eBay. Therefore, it is costly to amass an army of friends to trade recommendations with. However, eBay does not weight feedback by the amount of money paid to eBay; perhaps this is not costly enough. Such a system would require all traders to trust eBay's reporting - a decentralized peer to peer version is out of the question.

**Regret** is vulnerable to Sybil attack. It first finds the set of people who are connected to both the truster and trustee. It then selects a set of witnesses by finding cut-vertices and nodes with large degree of the components of this subgraph, both of which can be manipulated by adding fake identities. Then witnesses' recommendations are weighted by unspecified 'common sense' factors<sup>3</sup> or local trust values. The latter is only available if the witness and truster have interacted before.

**Mui** provides no mechanism for evaluating action histories. If they are blindly accepted, then attackers can falsely report arbitrarily many cooperation actions achieving arbitrarily high reputation scores.

**Aberrer** depends on all but a small portion of the population being honest. Therefore, it is vulnerable to Sybil attack.

Despite my decision not to include **collaborative filters** as reputation systems, it is worth noting that they are susceptible to Sybil attack. Consider a movie producer attempting to promote his movie on MovieLens. He, she, them and it could register different movie tastes, all including the producer's new flop.

## 2.4 The Need for Complaints

Reputation systems without support for complaints allow attackers to bait a community with good behaviour, and then exploit their good reputation indefinitely. The only way reputation can be lost is through withdrawals of recommendations. If an attacker avoids deceiving those who recommend him, then a system without complaints never provides an opportunity for bad behaviour to be reported back to the recommenders. Recommendation withdrawal would never happen.

Therefore, complaints should be an essential component of any reputation system. Of the systems surveyed above, only eBay and Aberrer have complaints. For example, Google's lack of complaints has resulted in a market for trading recommendations. Companies such as SearchKing<sup>4</sup> matchmake those who have accumulated a high PageRank score with those who want to increase their PageRank. While the market for hyperlinks may still exist within a system that supports complaints, it probably would not be as widespread because complaints would be registered against both the seller and buyer of reputation.

## 2.5 Problems with Complaints

Reputation Systems such as eBay that implement complaints have two extra problems to worry about. Firstly, after an honest trade, one party might threaten to complain about the other unless a 'certain payment' is made. [Resnick et al., 2000] Another problem is that a deceiving party can complain about an honest party's behaviour, in an attempt to discredit any complaints. It has been argued in [Resnick and Zeckhauser, 2002] and [Calkins, 2001] that complaints are under represented on eBay. [Calkins, 2001] conjectures that this might be because of widespread fear of a tit-for-tat retaliation.

## 2.6 Scope of Reputation

An important practical issue is dealing with the scope of recommendations and deciding what exactly a person has a good or bad reputation for. Consider the assertion 'Andrew is good at gardening'. What is gardening?

---

<sup>3</sup>They plan to do further research in this area

<sup>4</sup><http://www.searchking.com>

Does Andrew have a good reputation for recommending good florists? Are good florist recommenders good at recommending interior designers? These are philosophical issues that are investigated by *ontology* (the study of 'is').

[Sabater and Sierra, 2002] argues that all recommendations should be accompanied with a precise machine understandable description of the scope of the recommendation. Then, only the relevant recommendations need to be considered when computing a reputation score in a particular domain.

[Haveliwala, 2002] has suggested that Google could allocate the initial PageRank votes differently for each search query according to the topic of the query. For example, the initial votes could be configured to be a set of reputable law web sites when searching on a legal topic. [Haveliwala, 2002] considered using the Open Directory project categories for providing a suitable list of web pages for each topic, and showed how multiple categories can be combined efficiently.

## 2.7 Conclusion

Reputation is a convenient way of establishing trust with complete strangers. Reputation is useful for establishing trust both because past behaviour is indicative of future behaviour and because it introduces accountability. However, complaints are essential for this accountability. Moreover, weighting of recommendations by reputation scores is highly desirable for resisting Sybil attack. It has been conjectured that PageRank resists Sybil attack for this reason, but it lacks support for complaints. While eBay supports complaints, it lacks PageRank's recommendation weighting and is somewhat susceptible to Sybil attack.

## Chapter 3

# Defining PageRank

PageRank [Page et al., 1998] was developed by the founders of the popular Google [Brin and Page, 1998] search engine. PageRank is used by the search engine to assign a reputation score to every web page on the internet based on hyperlink structure and a set of initial votes obtained from a source such as the domain name website registration system. PageRank is important because it weights recommendations by the linking websites' reputation. This attractive feature has led to proposals to apply PageRank in eCommerce reputation [Yolum and Singh, 2003] and peer-to-peer networks [Kamvar et al., 2003].

While PageRank can be defined purely in terms of Linear Algebra, the theory of Markov processes is better developed for my purposes and provides richer intuition.

In 3.1 I review Markov theory to define PageRank and to provide background for later chapters. In 3.2, I define PageRank in terms of the Random Surfer model using Markov Theory. In 3.3, I resolve the confusion in the PageRank literature about the several inconsistent definitions of PageRank.

### 3.1 Markov Theory

This section introduces Markov theory in order to fix notation and provide background for the following chapters.

#### 3.1.1 Markov processes

The following definition defines a Markov process, the stochastic model used to model random surfers. In essence, the probability distribution of the next state (or next web page to visit) only depends on the current state (or current web page being visited). That is, the states prior to the current state are irrelevant (the web pages visited prior to the current web pages being visited are irrelevant). Also, the length of time for which the surfer has been surfing is irrelevant.

**Definition 3.1 (Markov process).** *An  $P$ -valued **Markov process** is an infinite sequence of random variables  $\{X_k\} = X_0, X_1, \dots \in P$  if  $P$  is finite and the probability function  $\mathbf{P}$  satisfies:*

$$\mathbf{P}(X_{k+1} = b \mid X_0 = a_0, \dots, X_k = a_k) = \mathbf{P}(X_{k+1} = b \mid X_k = a_k) \quad \text{is the same for all } k \geq 0$$

*Its **transition function** is  $\omega(i, j) = \mathbf{P}(X_{k+1} = b \mid X_k = a)$ . Its **initial distribution** is  $\sigma(a) = \mathbf{P}(X_0 = a)$*

In the Stochastic processes literature, this is technically called a *homogeneous, discrete time, finite space Markov process*. In applications of the theory, they are often simply called Markov processes or Markov chains.

### 3.1.2 Convergence of Markov processes

In this section, I review the conditions under which  $\lim_{k \rightarrow \infty} \mathbf{P}(X_k = a)$  converges.

Most of the Markov processes I will be discussing have a nice property called *ergodicity*. To define this, I need to define the *period of a state* first. Intuitively, if the only way from a state back to itself is through a cycle, then that state is periodic. If every state has the same period, then everything moves ‘in sync’, affecting its convergence properties.

**Definition 3.2 (Period of a state).** Let  $\{X_k\}$  be an  $P$ -valued Markov process. The **period** of a state  $p \in P$  is the largest  $d$  satisfying: (for all  $k, n \in \mathbf{N}$ )

$$\mathbf{P}(X_{k+n} = p | X_k = p) > 0 \implies d \text{ divides } n$$

If  $d = 1$ , then the state  $p$  is **aperiodic**.

For example, in a directed 2-cycle with transitions made with probability 1, both states have a period of 2.

**Definition 3.3 (Ergodic Markov process).** An **ergodic** Markov process is a Markov process  $\{X_k\}$  that is both:

- **irreducible**: every state is reachable from every other state.
- **aperiodic**: the greatest common divisor of the states’ periods is 1.

This following lemma gives us a simpler condition for ergodicity than verifying aperiodicity directly.

**Lemma 3.4 (Ergodic Condition).** An irreducible  $P$ -valued Markov process with transition function  $\omega$  that has  $\omega(a, a) > 0$  for some state  $a \in P$  is aperiodic, and hence ergodic.

*Proof.* This is a standard result. Firstly,  $\omega(a, a) = \mathbf{P}(X_{k+1} = a | X_k = a) > 0$ . So, the period of  $a$  must divide 1. Therefore, the period of  $a$  is 1 and  $a$  is aperiodic. The gcd of 1 and any other set of numbers is 1, so the Markov process is aperiodic, and hence ergodic.  $\square$

The following is a major theorem from Markov theory. Interested readers can read [Kemeny and Snell, 1976] for more information about ergodic convergence.

**Theorem 3.5 (Ergodic Convergence).** If  $\{X_k\}$  is an ergodic  $P$ -valued Markov process, then the probability function converges for all  $a \in P$ :

$$\lim_{k \rightarrow \infty} \mathbf{P}(X_k = a) = p_a$$

*Proof.* This proof is quite involved. It is the main topic of part I of [Behrends, 1999], for example.  $\square$

This following lemma allows us to understand two-state Markov processes (or if you prefer, a random surfer on a world wide web consisting of two web pages).

**Corollary 3.6 (Two-state Convergence).** Let the random variables  $\{X_k\}$  be a Markov process with states  $P = \{a, b\}$  and a transition function  $\omega$  such that  $\omega(a, b), \omega(b, a) > 0$  and either  $\omega(a, a) > 0$  or  $\omega(b, b) > 0$ , then:

$$\lim_{k \rightarrow \infty} \mathbf{P}(X_k = a) = \frac{\omega(b, a)}{\omega(a, b) + \omega(b, a)}$$

*Proof.* This is a standard result, and is a running example in [Norris, 1997].

By *Lemma 3.4 (Ergodic Condition)*,  $\{X_k\}$  is an ergodic Markov process. By *Theorem 3.5 (Ergodic Convergence)*, we know that  $\lim_{k \rightarrow \infty} \mathbf{P}(X_k = a)$  converges.

Now, I only need to deduce what  $\lim_{k \rightarrow \infty} \mathbf{P}(X_k = a)$  converges to.

From the *Definition 3.1 (Markov Process)*, we can write  $\mathbf{P}(X_{k+1} = a)$  in terms of  $\mathbf{P}(X_k = a)$ :

$$\begin{aligned} \mathbf{P}(X_{k+1} = a) &= \mathbf{P}(X_k = a) \cdot \omega(a, a) + \mathbf{P}(X_k = b) \cdot \omega(b, a) \\ &= \mathbf{P}(X_k = a) \cdot (1 - \omega(a, b)) + (1 - \mathbf{P}(X_k = a)) \cdot \omega(b, a) \\ &= \mathbf{P}(X_k = a) \cdot (1 - \omega(a, b) - \omega(b, a)) + \omega(b, a) \end{aligned}$$

We can rewrite this and find the limits to obtain the result:

$$\begin{aligned} \mathbf{P}(X_{k+1} = a) - \mathbf{P}(X_k = a) + \mathbf{P}(X_k = a)(\omega(a, b) + \omega(b, a)) &= \omega(b, a) \\ \lim_{k \rightarrow \infty} \mathbf{P}(X_{k+1} = a) - \lim_{k \rightarrow \infty} \mathbf{P}(X_k = a) + \lim_{k \rightarrow \infty} \mathbf{P}(X_k = a)(\omega(a, b) + \omega(b, a)) &= \lim_{k \rightarrow \infty} \omega(b, a) \\ \lim_{k \rightarrow \infty} \mathbf{P}(X_k = a)(\omega(a, b) + \omega(b, a)) &= \omega(b, a) \\ \lim_{k \rightarrow \infty} \mathbf{P}(X_k = a) &= \frac{\omega(b, a)}{\omega(a, b) + \omega(b, a)} \end{aligned}$$

□

### 3.1.3 Markov Process Isomorphisms

Here, I introduce Markov process isomorphisms, that allow us to determine whether two Markov processes have the same properties or not.

I have not been able to find any usage of ‘Markov process isomorphisms’ in the literature. This is probably because neither pure nor applied mathematicians have any need to use this concept. Applied mathematicians usually only compare Markov processes that share the same state space, in which case direct equality is sufficient. Unfortunately, this will not be the case for my purposes. Pure mathematicians would usually construct isomorphisms on the underlying measure space (i.e. bijections on the  $\sigma$ -algebras that preserve the measure). This level of generality is unnecessary in this thesis.

**Definition 3.7.** Let  $\{X_k\}$  be a  $P$ -valued Markov process. Let  $\{Y_k\}$  be a  $Q$ -valued Markov process.

Then a bijective function  $\xi : P \rightarrow Q$  is a **Markov process isomorphism** if:

$$\mathbf{P}(X_k = a) = \mathbf{P}(Y_k = \xi(a)) \quad \text{for all } k \in \mathbf{N} \text{ and all } a \in P \quad (3.1)$$

If an isomorphism between  $\{X_k\}$  and  $\{Y_k\}$ , then I will write:

$$\{X_k\} \cong \{Y_k\} \quad (3.2)$$

Obviously, since a Markov process is uniquely defined by its initial distribution and transition function, preservation of initial distribution and transition function of a bijection is both a necessary and sufficient condition for Markov isomorphism.

**Theorem 3.8 (Markov Isomorphism Condition).** Let  $\{X_k\}$  be a  $P$ -valued Markov process with transition function  $\omega$  and initial distribution  $\sigma$ . Let  $\{Y_k\}$  be a  $Q$ -valued Markov processes with transition function  $\psi$ , and initial distribution  $\tau$ .

Then a bijective function  $\xi : P \rightarrow Q$  is a Markov process isomorphism if and only if  $\sigma(a) = \tau(\xi(a))$  and  $\sigma(a, b) = \tau(\xi(a), \xi(b))$  for all  $a, b \in P$ .

*Proof.* Trivial: use induction on  $k$ . □

### 3.1.4 Lumpable Markov processes

This section reviews how states can be lumped together to form new Markov processes.

A partition of a state space is just a carving-up of a state-space into chunks.

**Definition 3.9 (Partition).** A *partition*  $W = \{W_0, W_1, \dots, W_n\}$  of a Markov process' state space  $P$  is a disjoint set of non-empty subsets of  $W$  such that

$$\cup_{i=0}^n W_i = P$$

The following defines a lumpable Markov process as a combination of a Markov process and partitioning that gives rise to a smaller stochastic process that is Markov. Later, I will use this to consider the lumped Markov process in which the entire world wide web is partitioned into two sets: pages belonging to the attacker and the rest of the web. Each random variable in this lumped process is the partition in which the page the Random Surfer is visiting. The lumped Markov process allows us to compute the sum of all PageRanks in each partition.

Note that this is called *weak lumpability*, because the term *lumpability* is reserved for Markov processes that have this property regardless of their starting distributions, which I do not consider in this thesis.

**Definition 3.10 (Weakly lumpable Markov process).** Let  $\{X_k\}$  be an  $P$ -valued Markov process. Let  $W = \{W_0, W_1, \dots, W_n\}$  be a partition of  $P$ . Let  $\rho : P \rightarrow W$  be a function such that  $\rho(w) = W_i$  when  $w \in W_i$ . Then  $\{X_k\}$  is **weakly lumpable** with respect to  $W$  if  $\{Y_k\} = \{\rho(X_0), \rho(X_1), \dots\}$  is a Markov process.

This theorem gives the conditions for which the lumped stochastic process is a Markov process: each state in a partition (lump) must share the same probability distribution of jumping to other partitions.

**Theorem 3.11 (Weakly lumpable condition).** Let  $\{X_k\}$  be an  $P$ -valued Markov process with transition function  $\omega$ . Let  $W = \{W_0, W_1, \dots, W_n\}$  be a partition of  $P$ . Let  $\rho : P \rightarrow W$  be a function such that  $\rho(w) = W_i$  when  $w \in W_i$ .

$\{X_k\}$  is weakly lumpable if and only if the function  $\phi : W \times W \rightarrow [0, 1]$  is well defined (unique):

$$\phi(W_i, W_j) = \sum_{b \in W_j} \omega(a, b) \quad \text{for all } a \in W_i$$

Furthermore, if  $\{X_k\}$  is lumpable, then  $\phi$  is the transition function of the Markov process  $\{Y_k\} = \{\rho(X_k)\}$ .

*Proof.* This is a standard result which appears in [Kemeny and Snell, 1976]. I will only show that it is a sufficient condition.

The general idea is, that if the transition functions are uniform with respect to partitions, then knowing information in addition to which partition  $X_k$  lies in does not affect the transition probabilities, and the Markov property is therefore inherited.

Assume that  $\phi$  is well-defined. Then:

$$\begin{aligned} & \mathbf{P}(\rho(X_{k+1}) = W_{i_{k+1}} \mid \rho(X_k) = W_{i_k}) \\ &= \mathbf{P}(X_{k+1} \in W_{i_{k+1}} \mid X_k \in W_{i_k}) \\ &= \sum_{b \in W_{i_{k+1}}} \mathbf{P}(X_{k+1} = b \mid X_k \in W_{i_k}) \\ &= \phi(\rho(a_k), W_{i_k}) && \text{for all } a_k \in W_{i_k} \\ &= \mathbf{P}(X_{k+1} \in W_{i_{k+1}} \mid X_k = a_k) && \text{for all } a_k \in W_{i_k} \\ &= \mathbf{P}(X_{k+1} \in W_{i_{k+1}} \mid X_0 = a_0, \dots, X_k = a_k) && \text{for all } a_k \in W_{i_k} \\ &= \mathbf{P}(X_{k+1} \in W_{i_{k+1}} \mid X_0 \in W_{i_0}, \dots, X_k \in W_{i_k}) \\ &= \mathbf{P}(\rho(X_{k+1}) = W_{i_{k+1}} \mid \rho(X_0) = W_{i_0}, \dots, \rho(X_k) = W_{i_k}) \end{aligned}$$

This final equality is exactly the definition of a Markov process. So  $\{\rho(X_k)\}$  is a Markov process, and  $\{X_k\}$  is lumpable with respect to  $\{W_0, \dots, W_j\}$ . Furthermore,  $\phi$  is the transition function.  $\square$

For convenience, I will use the following division notation for constructing lumped Markov processes:

**Definition 3.12 (Division Notation for Lumping).** *Let  $\{X_k\}$  be an  $P$ -valued Markov process. If  $W \subset P$ , then  $\{Y_k\} = \{X_k\}/W$  is the  $P'$ -valued stochastic process with  $Y_k = \rho(X_k)$  where:*

$$P' = (P \setminus W) \cup \{W\} \tag{3.3}$$

$$\rho(a) = \begin{cases} a & \text{if } a \in P \setminus W \\ W & \text{if } a \in W \end{cases} \tag{3.4}$$

### 3.1.5 Summary of Markov Theory

This concludes the material I will draw on from Markov theory. In summary, a (discrete time homogeneous) Markov process is a stochastic process where has a well-defined transition probability function that only depends on the current state. Well-behaved ergodic Markov processes have their probabilities converge to an equilibrium. We can easily compute this equilibrium for two-state Markov processes. Moreover, under certain conditions, Markov processes can be collapsed by lumping states together.

## 3.2 Random Surfer Model

In this section, I will define PageRank in terms of the Random Surfer model. I will first describe the intuition behind PageRank and then formalize it.

### 3.2.1 Intuition

Perhaps the best measure of web page quality would be the total amount of time all web surfers spend visiting a particular web page as a fraction of total web surfing time. Unfortunately, this information is not available. Even if web surfers did not object to having their movements watched, it would be difficult to verify whether a web surfer was really visting a page. The next best thing then is to model the behaviour of web surfers and estimate this quantity. This is the essence of PageRank.

PageRank's Random Surfer model supposes a random surfer that can only click on hyperlinks and jump to a random web page. That is, at each time step, there is a chance that the random surfer gets bored and rather than click on a hyperlink, visits an imaginary random search engine that delivers them to a page randomly. The random surfer does not have a 'back button'.

The PageRank of a web page then is the proportion of time the random surfer would spend at that web page if it continued to surf forever.

Under this model, PageRank is parameterized by the probabilities in the random search engine for delivering a surfer to a particular page and the probability that the random surfer gets bored. The former controls the initial votes and the later the amount reputation is allowed to flow from these initial votes.

### 3.2.2 Formal definition

I will now formally define PageRank in terms of the Random Surfer model.

I will first define a simple *Random Click* model. This will lack ergodicity and does not provide any mechanism for seeding the reputation flow. I will then define the Random Surfer model, which I will call the *Random Search-Click* model to guide intuition, that addresses both of these problems. I will define the PageRank of a webpage to be the equilibrium probability of a random surfer following this model. I will show that this definition of PageRank is well-defined.

First I will define a webpage, a model of the structure of the world wide web:

**Definition 3.13 (Webgraph).** Let  $G = (P, H)$  be a finite directed graph, where  $P$  is the set of web pages, and  $H \subseteq P \times P$  the set of hyperlinks.  $G$  is **webgraph** if all pages have an outgoing hyperlink (possibly to themselves).

I will now define the *Random Click* process. The random variables  $\{X_k\}$  refer to the webpages (elements of  $P$ ) that the random surfer visits at each timestep. The surfer either clicks on a random hyperlink, or does nothing at each time step. There are no other possibilities (such as jumping to a home page or search engine). Since the surfer must choose one of these possibilities, a web page that has no outgoing hyperlinks will force the surfer to ‘choose’ to do nothing.

The only assumption I make about the probability distribution of clicking on a particular link is that there is a non-zero chance of following any link of a page the surfer is visiting.

**Definition 3.14 (Random Click process).** A **Random Click process** of a webgraph  $G = (P, H)$  is a  $P$ -valued Markov process  $\{X_k\}$  such that transition function  $\omega$  has  $\omega(a, b) > 0$  if<sup>1</sup> and only if  $(a, b) \in H$ .

In general, Random Click processes are not ergodic.

I now add a random search engine to *Random Click* processes to form *Random Search-Click* processes. If the random surfer chooses to use the random search engine rather than following a hyperlink, it is delivered to a random web page according with the probability distribution defined by the Random Search function.

This addition makes *Random Search-Click* processes almost<sup>2</sup> ergodic. The Random Search function also allows us to place *a priori* value judgements on web pages.

**Definition 3.15 (Random Search function).** A **Random Search function** of a webgraph  $G = (P, H)$  is a function  $s : P \rightarrow [0, 1]$  with:

$$s(P) := \sum_{p \in P} s(p) = 1$$

One example of a Random Search function is a uniform distribution:

$$s(p) = \frac{1}{|P|}$$

Now, I will define the Random Search-Click model. A Random Search-click process is constructed from a webgraph, a click distribution, a search probability  $d$  and a Random Search function. At each time step, the random surfer chooses to use the random search engine with (constant) probability  $d$ , or to follow a hyperlink (as in Random Click processes) with probability  $1 - d$ . [Page et al., 1998] suggests 0.15 as an appropriate value for  $d$ .<sup>3</sup> Since the surfer’s choice of click vs search is geometrically distributed, this value gives the random surfer an average of  $\frac{1}{0.15} - 1 \approx 6$  clicks before doing a random search.

This addition of a random search engine prevents the random surfer from ever getting stuck in a dead end or closed cycle. This makes Random Search-Click processes close enough to being ergodic that an equilibrium distribution exists.

The Random Search function also provides a mechanism for allocation of initial votes.

Again, the random variables  $\{X_k\}$  refer to the webpages (elements of  $P$ ) that the random surfer visits at each timestep.

**Definition 3.16 (Random Search-Click process).** Let  $G = (P, H)$  be a webgraph. Let  $s$  be a Random Search function of  $G$ . Let  $d \in (0, 1)$  be an arbitrary constant. Let  $\omega$  be the transition function of some Random Click process on  $G$ .

---

<sup>1</sup>This definition has a tight “if and only if” to simplify the presentation so “connectivity in  $G$ ” can be used synonymously with “reachable with nonzero probability”. In practice, this can be relaxed to “only if” by removing each edge  $ab$  from  $G$  where  $\omega(a, b) = 0$ .

<sup>2</sup>Possibly ergodic  $\mathbf{P}$ -almost-everywhere. I did not investigate this possibility. It could make the proofs more elegant.

<sup>3</sup>It is not quite clear what they intended this 0.15 value refer to. This is discussed in the last section of this chapter.

Then,  $\{X_k\}$  is a  $(G, \omega, d, s)$  Random Search-Click process if  $\{X_k\}$  is the  $P$ -valued Markov process with initial distribution  $s$  and transition function  $\psi : P \times P \rightarrow [0, 1]$ :

$$\begin{aligned}\mathbf{P}(X_0 = a) &= s(a) \\ \psi(a, b) &= ds(b) + (1 - d)\omega(a, b)\end{aligned}$$

I am now ready to define PageRank of a page as the equilibrium probability of that page in a particular Random Search-Click process:

**Definition 3.17 (PageRank).** Let  $\{X_k\}$  be a Random Search-Click process on a webgraph, with transition function  $\psi$ . The **PageRank**  $r(a)$  of a page  $a \in P$  is:

$$r(a) = \lim_{k \rightarrow \infty} \mathbf{P}(X_k = a)$$

The parameters for constructing this Random Search-Click process underlying PageRank are the webgraph  $G$ , the Random Search function  $s$ , the search probability  $d$ , and the click distribution  $\omega$ .

I need to show that every page has a well-defined PageRank value. To begin, I show that a webpage has a zero pagerank if and only if it is unreachable from a webpage in the random search engine:

**Lemma 3.18 (Zero Condition).** Let  $\{X_k\}$  be a  $(G, \omega, d, s)$  Random Search-Click process. The PageRank of a page  $a$  has  $r(a) = 0$  if and only if there is no  $ba$ -path (in  $G$ ) from some  $b \in P$  with  $s(b) > 0$ .

*Proof.* Firstly, assume that  $r(a) = 0$  (that is, that it converges and it converges to this particular value). Now assume for the sake of contradiction that there exists some  $ba$ -path with pages  $p_0 = b, p_1, \dots, p_l = a$  having  $s(b) > 0$ . Clearly, for all  $k$ ,  $\mathbf{P}(X_k = b) \geq s(b)$ . So:

$$\begin{aligned}\mathbf{P}(X_{k+l} = a) &\geq \mathbf{P}(X_k = b) \cdot \mathbf{P}(X_{k+l} = a \mid X_k = b) \\ &= \mathbf{P}(X_k = b) \cdot \prod_{i=0}^l \mathbf{P}(X_{k+i+1} = p_{i+1} \mid X_{k+i} = p_i) \\ &= \mathbf{P}(X_k = b) \cdot c && \text{where } c > 0 \text{ is a constant independent of } k \\ &\geq s(b) \cdot c \\ &> 0\end{aligned}$$

Now, consider the value of  $r(a)$ :

$$\begin{aligned}r(a) &= \lim_{k \rightarrow \infty} \mathbf{P}(X_k = a) \\ &= \lim_{k \rightarrow \infty} \mathbf{P}(X_{k+l} = a) \\ &\geq s(b) \cdot c \\ &> 0\end{aligned}$$

But I assumed that  $r(a) = 0$ , so there is a contradiction. Therefore, the assumption that there exists such a  $ba$ -path is false, and no such path exists.

Conversely, assume that no  $ba$ -path exists for any  $b$  with  $s(b) > 0$ . Then  $\mathbf{P}(X_k = a) = 0$  for every  $k$ . (If this were not the case, then  $X_0, \dots, X_k$  would contain such a  $ba$ -path). It follows, that  $r(a) = \lim_{k \rightarrow \infty} \mathbf{P}(X_k = a) = 0$  as required.  $\square$

I will now show that this is well defined. This is well known, but a proof has not appeared in this general form.

**Theorem 3.19 (PageRank well-defined).** Let  $\{X_k\}$  be a  $(G, \omega, d, s)$  Random Search-Click process. The PageRanks  $r(a)$  for all web pages  $a \in P$  are well defined.

*Proof.* To show that  $r(a)$  is well defined, I need to show that  $r(a)$  exists and is unique. I will show that removing all web pages that are unreachable from a sequence of clicks beginning at the random search engine yields an ergodic Markov process. Then, I will be able to invoke *Theorem 3.5 (Ergodic Convergence)* to show that  $r$  exists and is unique.

Let  $P' = \{a \in P \mid \mathbf{P}(X_k = a) > 0 \text{ for some } k\}$ , the reachable pages in  $P$ . Then  $\{X_k\}$  is a  $P'$ -valued Markov process. Clearly,  $\{X_k\}$  is irreducible as a  $P'$ -valued Markov process, because every state is reachable from every other state via some state  $b$  with  $s(b) > 0$ .

Also, there exists a page  $a \in P'$  with  $s(a) > 0$ , and hence  $\omega(a, a) > 0$ . Again, by *Lemma 3.4*,  $X_0, X_1, \dots$  is ergodic as a  $P'$ -valued Markov process.

By *Theorem 3.5 (Ergodic Convergence)*, a unique  $r(a)$  value exists for every page  $a \in P'$ . By *Lemma 3.18 (Zero condition)*,  $r(a) = 0$  for every page in  $P \setminus P'$ .  $\square$

This completes my definition of PageRank in terms of the Random Surfer model.

### 3.2.3 Algebraic PageRank

It turns out that PageRank weights recommendations by the recommender's reputation. In fact, this property uniquely defines PageRank as well. It is commonly used as the definition of PageRank (for example in [Haveliwala, 2002]).

First, I will define some notation:

**Definition 3.20 (Neighbourhoods).** *If  $G = (P, H)$  is a webgraph, then define the incoming and outgoing neighbourhoods of  $a$  as:*

$$N^-(a) = \{b : (b, a) \in H\}$$

$$N^+(a) = \{b : (a, b) \in H\}$$

This theorem shows that PageRank weights recommendations:

**Theorem 3.21 (PageRank Algebra).** *Let  $\{X_k\}$  be a  $(G, \omega, d, s)$  Random Search-Click process. Then, for all  $a \in P(G)$ :*

$$r(a) = ds(a) + (1 - d) \sum_{b \in N^-(a)} \omega(b, a)r(b)$$

*Proof.* From *Definition 3.17 (PageRank)*, we have:

$$\begin{aligned} r(a) &= \lim_{k \rightarrow \infty} \mathbf{P}(X_k = a) \\ &= \lim_{k \rightarrow \infty} \sum_{b \in P} \mathbf{P}(X_k = a \mid X_{k-1} = b) \mathbf{P}(X_{k-1} = b) \\ &= \sum_{b \in P} \lim_{k \rightarrow \infty} \psi(b, a) \mathbf{P}(X_{k-1} = b) \\ &= \sum_{b \in P} \psi(b, a) r(b) \\ &= \sum_{b \in P} [ds(a) + (1 - d)\omega(b, a)] r(b) \\ &= ds(a) \sum_{b \in P} r(b) + (1 - d) \sum_{b \in P} \omega(b, a) r(b) \\ &= ds(a) + (1 - d) \sum_{b \in N^-(a)} \omega(b, a) r(b) \end{aligned}$$

$\square$

Name	Appearances	Proofs	Comments
<i>PageRank</i>	[Haveliwala, 2002, Kamvar et al., 2003]	<i>Theorem 3.19</i>	most general
<i>PageRank1</i>	[Bianchini et al., 2004]	<i>Corollary 3.25</i>	uniform votes
<i>PageRank1Wrong</i>	[Brin and Page, 1998]	<i>Claim 3.24</i>	sum $\neq 1$
<i>PageRank2</i>	[Page et al., 1998]	<i>Corollary 3.27</i>	unintuitive

Table 3.1: The different definitions of PageRank

### 3.3 Definition Confusion and Contradiction

There is considerable confusion about the definition of PageRank. Several different but equivalent definitions of PageRank are available. Even the two papers ([Brin and Page, 1998] and [Page et al., 1998]) by its developers use different definitions without any explanation. [Bianchini et al., 2004] noticed the two definitions were inconsistent with each other, but that they are still closely related. However, the first definition is inconsistent with a statement made in the first paper. I argue that the definition widely used is what the authors really intended.

To resolve these issues, I will present the two definitions from the two papers by the PageRank developers, and argue that the first of these is inconsistent, and suggest what they really had in mind. Then, I will show that the corrected PageRank definition of the first paper and the definition of the second paper are special cases of my definition.

To distinguish between the various definitions of PageRank, I will call the PageRank as defined in this paper *PageRank* with notation  $r(a)$ , the PageRank in [Brin and Page, 1998] *PageRank1Wrong* with notation  $r_{1w}(a)$ , the PageRank that I think [Brin and Page, 1998] intended to define as *PageRank1* with notation  $r_1(a)$  and the PageRank in [Page et al., 1998] *PageRank2* with notation  $r_2(a)$ .

#### 3.3.1 *PageRank1Wrong* as defined in [Brin and Page, 1998]

Unfortunately, there is a mistake that I discovered in the presentation of *PageRank1Wrong* in [Brin and Page, 1998]. The authors omitted a division, that in my terminology means that the random search function summed to  $|P|$  rather than 1. Under this mistaken definition, the PageRanks do not represent probabilities as they do not sum to 1. Most researchers (including [Haveliwala, 2002]) use a different definition that is consistent with mine, perhaps noting that it is different (as in [Bianchini et al., 2004]) without labelling it a mistake.

In this section, I will argue that there is indeed a mistake and describe what I believe the authors really had in mind. I will prove that this corrected version, *PageRank1*, is a special case of *PageRank* as I defined it in *Definition 3.17*.

Here it was defined incorrectly:

**Definition 3.22 (PageRank1Wrong).** *If  $G = (P, H)$  is a webgraph, then the PageRank1  $r_{1w}(a)$  of a page  $a$  is:*

$$r_{1w}(a) = (1 - c) + c \sum_{b \in N^-(a)} \frac{r_{1w}(b)}{|N^+(b)|}$$

where  $c$  is a constant parameter.

It should have been defined like this:

**Definition 3.23 (PageRank1).** *If  $G = (P, H)$  is a webgraph, then the PageRank1  $r_1(a)$  of a page  $a$  is:*

$$r_1(a) = \frac{1 - c}{|P|} + c \sum_{b \in N^-(a)} \frac{r_1(b)}{|N^+(b)|}$$

where  $c$  is a constant parameter.

[Brin and Page, 1998] proposes that a value of 0.85 for  $c$  is a good choice. Their constant  $c$  is equivalent to  $1 - \frac{E}{E+1}$  presented in [Page et al., 1998], and  $1 - d$  presented in this paper in *Definition 3.16 (Random Search-Click process)*. We chose to be closer to [Page et al., 1998] in this respect. This value is called the *damping factor* in [Brin and Page, 1998] and the *decay factor* in [Page et al., 1998].

To verify that their original definition is indeed mistaken, I quote from their paper [Brin and Page, 1998] this observation:

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

I claim that *PageRank1Wrong* does not have this property, but *PageRank1* does. I will prove that *PageRank1* has this property by showing that it is a special case of *PageRank* (as defined in *Definition 3.17*), and hence is a probability distribution that satisfies this property.

**Claim 3.24 (Incorrectness of PageRank1Wrong).** *Let  $G = (P, H)$  be a webgraph. Let  $r_{1w}(a)$  be the PageRank1Wrong score of webpage  $a$ , with parameter  $c > 0$ . Then  $r_{1w}(a)$  does **not** always form a probability distribution, such that  $r_{1w}(a) \geq 0$  for all  $a \in P$  and  $\sum_{a \in P} r_{1w}(a) = 1$ .*

*Proof.* For the sake of contradiction, assume that  $r_{1w}(a) \geq 0$  for all  $a \in P$  and  $\sum_{a \in P} r_{1w}(a) = 1$ .

Recall the definition of  $r_{1w}$ :

$$r_{1w}(a) = (1 - c) + c \sum_{b \in N^-(a)} \frac{r_{1w}(b)}{|N^+(b)|}$$

Now, observe that as  $c > 0$  and  $r_{1w} > 0$ , the second term is greater than zero, and hence  $r_{1w}(a) \geq 1 - c$ . Thus:

$$\begin{aligned} \sum_{a \in P} r_{1w}(a) &\geq \sum_{a \in P} (1 - c) \\ &\geq |P|(1 - c) \end{aligned}$$

Since  $1 - c$  is a fixed parameter, and  $|P|$  can grow arbitrarily, I conclude that this does not always equal 1. Therefore, the claim that *PageRank1Wrong* always forms a probability distribution is incorrect.  $\square$

In practice, this leaves a large discrepancy. For the suggested value of  $c = 0.85$ , and the reported 3 billion documents on the internet, the sum of PageRanks would be about 50 million. It is inconceivable that the authors of this paper intended to define anything other than *PageRank1*.

Now I will show that *PageRank1* is the special case of my *PageRank* where the Random Search function is uniform across all web pages, and the Random Surfer chooses hyperlinks with uniform probability. Note that [Bianchini et al., 2004] proved this theorem in a more complicated way using dynamical systems to attain a deeper result about how *PageRank1* is related to *PageRank1Wrong*.

**Corollary 3.25 (PageRank1 is a Special Case).** *Let  $\{X_k\}$  be a  $(G, \omega, s, d)$  Random Click process where the click distribution is  $\omega(a, b) = \frac{1}{|N^+(a)|}$  when  $(a, b) \in H$  and  $\omega(a, b) = 0$  otherwise, and  $s(a) = \frac{1}{|P|}$  be the uniform Random Search function.*

*If  $r$  is the PageRank function constructed from  $\{X_k\}$  and  $r_1$  is constructed from  $G$  and  $c = 1 - d$ , then  $r(a) = r_1(a)$  for all  $a \in P$ .*

*Proof.* From *Theorem 3.21 (PageRank Algebra)*, we have:

$$\begin{aligned}
r(a) &= ds(a) + (1-d) \sum_{b \in N^-(a)} \omega(b,a)r(b) \\
&= d \frac{1}{|P|} + (1-d) \sum_{b \in N^-(a)} \frac{1}{|N^+(b)|} r(b) \\
&= \frac{1-c}{|P|} + c \sum_{b \in N^-(a)} \frac{r(b)}{|N^+(b)|}
\end{aligned}$$

This is exactly the definition of  $r_1$ . □

This concludes the discussion of the definition of *PageRank1* in [Brin and Page, 1998]. In summary: the definition given was clearly a mistake, but can be easily corrected. The corrected version is a special case of my generalized definition in terms of Random Search-Click processes.

### 3.3.2 *PageRank2* as defined in [Page et al., 1998]

In this section, I show *PageRank2* as defined in [Page et al., 1998] is a special case of PageRank as defined in this paper. However, I argue that their choice of parameter  $E$  is clumsy.

*PageRank2* is defined in [Page et al., 1998] as follows:

**Definition 3.26 (PageRank2).** Let  $G = (P, H)$  be a webgraph and  $e : P \rightarrow R^+$  be a source of rank. The *PageRank2*  $r_2(a)$  of a web page  $a$  is:

$$r_2(a) = ce(a) + c \sum_{b \in N^-(a)} \frac{r_2(b)}{|N^+(b)|}$$

where  $c$  is chosen such that  $\sum_{a \in P} r_2(a) = 1$ .

Again, [Page et al., 1998] recommend  $e$  be chosen such that  $\sum_{a \in P} e(a) = 0.15$  (which is  $1 - 0.85$ ). However, this is *not* the same as the Random Search probability,  $d$ . I believe this is a mistake on the part of the authors of [Page et al., 1998].

**Corollary 3.27 (PageRank2 is a Special Case).** Let  $G = (P, H)$  be a webgraph. Let  $e : P \rightarrow R^+$  be a function. Let  $E = \sum_{a \in P} e(a)$ .

Let  $\omega$  be the click distribution with  $\omega(a,b) = \frac{1}{|N^+(a)|}$  when  $(a,b) \in H$  and  $\omega(a,b) = 0$  otherwise. Let  $s = \frac{e(a)}{E}$  be the Random Search function that is a scaled version of  $e$ . Let  $d = \frac{1}{1+E}$ . Let  $\{X_k\}$  be a  $(G, \omega, d, s)$  Random Search-Click process.

If  $r$  is the PageRank function constructed from  $\{X_k\}$  and  $r_2$  is constructed from  $G$  and  $e$ , then  $r(a) = r_2(a)$  for all  $a \in P$ .

*Proof.* From *Theorem 3.21 (PageRank Algebra)*, we have:

$$\begin{aligned}
r(a) &= ds(a) + (1-d) \sum_{b \in N^-(a)} \omega(b,a)r(b) \\
&= \left( \frac{E}{E+1} \right) \left( \frac{e(a)}{E} \right) + \left( 1 - \frac{E}{E+1} \right) \sum_{b \in N^-(a)} \frac{1}{|N^+(b)|} r(b) \\
&= \frac{1}{E+1} e(a) + \frac{1}{E+1} \sum_{b \in N^-(a)} \frac{r(b)}{|N^+(b)|}
\end{aligned}$$

Substituting  $c = \frac{1}{E+1}$ , we obtain:

$$r(a) = ce(a) + c \sum_{b \in N^-(a)} \frac{r(b)}{|N^+(b)|}$$

This is exactly the definition of  $r_2$ .  $c$  is chosen appropriately, since the sum of  $r$  values is 1. □

The fractions  $s(a) = \frac{e(a)}{E}$  and  $c = \frac{1}{E+1}$  are artifacts of the definition of *PageRank2*. *PageRank*'s definition is cleaner mathematically and matches the Random Surfer model in an obvious way.

### 3.3.3 Summary

*PageRank1Wrong* as defined in [Brin and Page, 1998] had a mistake, leading it to be inconsistent with the claim that all PageRanks sum to 1. This can be easily corrected to *PageRank1*, which is a special case of *PageRank* where the Random Search function  $s$  is uniform and the click distribution  $\omega$  is uniform along each hyperlink for each page.

*PageRank2* as defined in [Page et al., 1998] is consistent, but is somewhat clumsy in that it does not connect cleanly with the Random Surfer model. *PageRank2* is a special case of *PageRank* where the click distribution  $\omega$  is uniform along each hyperlink for each page.

## Chapter 4

# PageRank Cost of Attack

Reputation metrics must balance the desire to exploit as much information as possible with the desire to avoid being misled by deceptive information. In the previous chapter, I described how PageRank weights all information by the reputation of its source. In this chapter, I consider the implications of this weighting to the cost of attacking PageRank to deceptively acquire reputation.

I first develop some mathematical theory describing how PageRank’s initial votes affect the system. Then I apply this theory to estimate the cost of attack of Google. I compare this with the price list of a company that sells Google PageRanks. Finally, I discuss the implications of widespread public knowledge of the cost of attack of PageRank.

### 4.1 PageRank Cost of Attack Theorem

In this section, I prove that PageRank resists a large class of attacks. I introduce the theorem, then provide the intuition behind the proof and finally present the proof.

I frequently talk about the portion of the internet that the attacker controls, and the portion the attacker does not control. I will call the former the *attacker’s pages* and the latter the *original pages*. I assume that the attacker can not modify the original pages.

Formally, I prove the following:

**Theorem 4.4 (PageRank Cost of Attack).** *Let  $G = (V \cup W, H)$  be a webgraph where  $V \cap W = \emptyset$ . Let  $\{X_k\}$  be a  $(G, \omega, d, s)$  Random Search-Click process. Let  $r$  be the PageRank function using the process  $\{X_k\}$ . If there are no hyperlinks from  $W$  to  $V$  (i.e.  $H \cap (W \times V) = \emptyset$ ), then:*

$$\sum_{v \in V} r(v) \leq \sum_{v \in V} s(v) \tag{4.1}$$

*Moreover, equality occurs if there are no hyperlinks from  $V$  to  $W$ .*

This means that if the original pages have no hyperlinks to the attacker’s pages, then the sum of the PageRanks of the attacker’s pages is less than or equal to the proportion of initial votes allocated to the attacker’s pages. That is, there is no way for an attacker to gain a higher PageRank sum by deceptively adding or modifying the attacker’s pages.

PageRank has several parameters: the graph structure  $G$ , the click distribution  $\omega$ , the search probability  $d$  and search distribution  $s$ . The theorem says that if there are no hyperlinks from pages in  $W$  (the web) to pages in  $V$  (the attacker), then the graph structure and click distribution inside  $V$ , and the search probability  $d$  are irrelevant to the sum of the PageRanks in  $V$ . Only the search distribution matters.

Ideally, I would like to be able to deal with cases where there are some hyperlinks from the web at large to an attacker’s pages. This limitation will be discussed later.

In 4.1.1, I provide an overview of the proof that follows in 4.1.2.

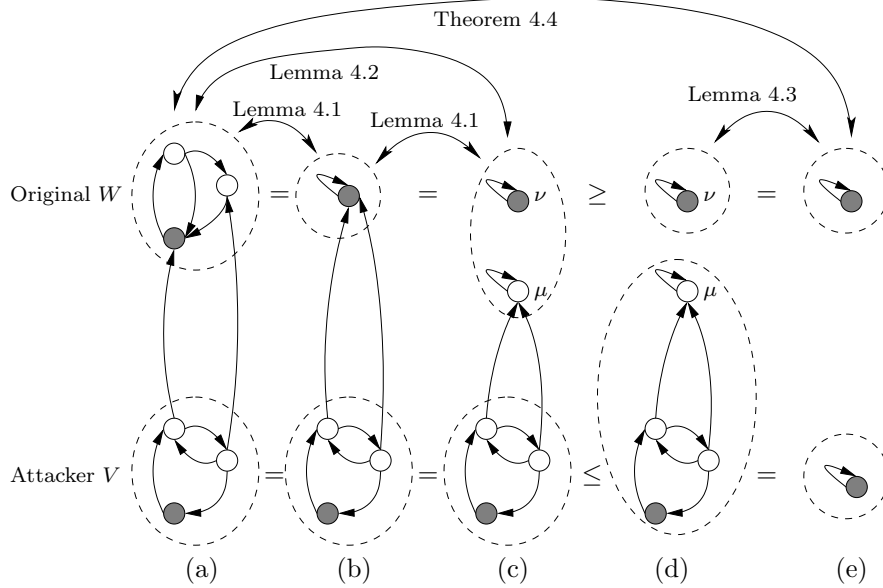


Figure 4.1: *Overview of the Cost of Attack Theorem.* The graph at the left represents an attacker trying to maximize his PageRank. Each successive graph has a successively simpler but equivalent graph. The right graph is simple enough to be able to compute PageRanks directly. The bottom represents web pages controlled by the attacker. The top represents pages not controlled by the attacker. The filled in vertices represent web pages that have been allocated a non-zero initial vote. The arcs represent hyperlinks. The dashed ellipses represent the sets of PageRanks I am computing the sum of at each step.

### 4.1.1 Overview

Here, I describe the intuition behind *Theorem 4.4 (PageRank Cost of Attack)*.

The Cost of Attack theorem is proved with the help of three lemmas. *Lemma 4.1 (PageRank Lumping)* describes how a set of pages can be lumped together into a single representative page in order to compute the sum of the set's PageRanks. *Lemma 4.2 (PageRank Lumping Isomorphism)* shows that the internal structure of the lumped pages in the previous lemma is irrelevant to the sum of the PageRanks, and can therefore be arranged arbitrarily. *Lemma 4.3 (PageRank Lump Sum)* is a special case of the main theorem. It gives the sum of PageRanks in a set of web pages where there are hyperlinks neither entering nor leaving the set.

#### Overview of Theorem 4.4 (PageRank Cost of Attack)

*Fig 4.1.1* shows how I put these lemmas together to prove *Theorem 4.4 (PageRank Cost of Attack)*. I reduce the general case (*Fig 4.1.1a*) where the attacker can add hyperlinks pointing to any web page to the special case where the attacker's hyperlinks can only point to the attacker's pages (*Fig 4.1.1d*). I deal with this special case in *Lemma 4.3 (PageRank Lump Sum)*.

I use *Lemma 4.2 (PageRank Lumping Isomorphism)* to split the original web pages into two representative web pages  $\mu$  and  $\nu$ .  $\mu$  represents the pages the attacker was stupid enough to link to. The PageRank of  $\mu$  is the amount of PageRank the attacker donated to the original pages.  $\nu$  represents the remaining original pages. The PageRank of  $\nu$  is the PageRank the original pages would have had if the attacker had not hyperlinked to them.

To do the reduction to *Fig 4.1.1d*, I lump  $\mu$  with the attacker's pages. Since the construction has no hyperlinks between  $\mu$  and  $\nu$ , there are no hyperlinks between this lump and  $\nu$ . Therefore, *Lemma 4.3*

(*PageRank Lump Sum*) can be applied to compute the sum of the PageRanks of the attacker’s pages and  $\mu$ . Clearly, this sum is greater than the sum of the PageRanks of the attacker’s pages alone.

### Overview of Lemma 4.1 (PageRank Lumping)

*Lemma 4.1 (PageRank Lumping)* is about shrinking a Random Search-Click process  $\{X_k\}$  by lumping some pages  $W$  together, and hence constructing a smaller Random Search-Click process  $\{X_k\}/W$ . The Lemma states that this is possible if there are no outgoing hyperlinks from the pages that are to be lumped together. Moreover, the sum of the PageRanks of the pages in the lump equals the PageRank of the lump in the lumped process.

To prove that the lumped process  $\{X_k\}/W$  is a Random Search-Click process (under the hyperlink condition), I first use *Theorem 3.11 (Lumpability Condition)* to prove it is a Markov process. Then, I construct a Random Search-Click process that shares the transition function with the lumped process.

The fact that the sum of the PageRanks in the lump  $W$  equals the PageRank of the lump is a trivial fact. Since ‘the Random Surfer is visiting pages  $a$  at time  $k$ ’ and ‘the Random Surfer is visiting page  $b$  at time  $k$ ’ cannot occur simultaneously, the probability of either happening is equal to the sum of the probabilities of each happening. PageRank, being the limit as  $k \rightarrow \infty$ , clearly shares this property.

### Overview of Lemma 4.2 (PageRank Lumping Isomorphism)

*Lemma 4.2 (PageRank Lumping Isomorphism)* gives a sufficient condition for two Random Search-Click processes  $\{X_k\}$  and  $\{Y_k\}$  to have isomorphic lumped processes  $\{X_k\}/W_X \cong \{Y_k\}/W_Y$ . For example, the lumped processes of *Fig 4.1.1a*, *Fig 4.1.1b* and *Fig 4.1.1c* are all isomorphic to each other, where the Original pages ( $W$ ) are lumped together.

The condition is that the two processes must share the same properties outside of  $W_X$  and  $W_Y$ , and there must not be any outgoing hyperlinks leaving  $W_X$  and  $W_Y$ . This condition does not constrain the internal structure (click and search distributions) of  $W_X$  and  $W_Y$ . This means isomorphic Random Search-Click processes with arbitrary structure inside the lump can be constructed. Equivalently, it means that lumping destroys the internal structure of the lump completely. Therefore, no matter what lump is supplied for destruction, the same resulting Random Search-Click process emerges.

Firstly,  $\{X_k\}/W_X$  and  $\{Y_k\}/W_Y$  are Random Search-Click processes, by the previous lemma.

Secondly, since each state’s click distribution must sum to 1, the probability of clicking in a link into the lump is simply 1 minus the sums of the probabilities on clicking on links outside the lump. This means the transition function for  $\{X_k\}/W_X$  for states outside of  $W_X$  is uniquely defined by the search and click distributions of  $\{X_k\}$  outside of  $W_X$ . Moreover, since no hyperlinks are allowed to leave the lump, the click distribution for the lumped page is trivially 1 for the self-link, and 0 for everything else.

Therefore, the transition function for  $\{X_k\}/W_X$  is uniquely for all states defined by the structure of  $\{X_k\}$  outside of  $W_X$ . Whenever this structure is shared with  $\{Y_k\}$  outside of  $W_Y$ , the transition function of  $\{Y_k\}/W_Y$  is also uniquely defined, and must therefore match that for  $\{X_k\}/W_X$ .

As a technical detail, the state spaces are different, so direct equality does not hold. Instead, I construct an isomorphism.

### Overview of Lemma 4.3 (PageRank Lump Sum)

*Lemma 4.3 (PageRank Lump Sum)* states that if the set of web pages can be bi-partitioned such that there are no hyperlinks between pages in different partitions, then the sum of the PageRanks in a partition equals the sum of the initial votes allocated to that partition. *Fig 4.1.1d* and *Fig 4.1.1e* are examples of where this Lemma may be applied.

There are two steps to the proof. First, the Random Search-Click process is lumped once for each partition. Then, the PageRanks in the doubly lumped process are computed using *Corollary 3.6 (Two-state Convergence)*. The algebra falls out nicely to get the result.

### 4.1.2 Mathematical Proof

Here I formalize the overview of the proof given above of *Theorem 4.4 (PageRank Cost of Attack)*.

This following Lemma is about grouping web pages together, and treating them as if they were a single web page. It says that you can do this if there are no hyperlinks from the group to outside of the group. Moreover, the PageRank of the single grouped web page equals the sum of all the PageRanks of the web pages in the group.

**Lemma 4.1 (PageRank Lumping).** *Let  $G = (V \cup W, H)$  where  $V \cap W = \emptyset$ . Let  $\{X_k\}$  be a  $(G, \omega, d, s)$  Random Search-Click process with transition function  $\psi$ . Let  $\{X'_k\} = \{X_k\}/W$ . Let  $r$  and  $r'$  be the PageRank functions using the processes  $\{X_k\}$  and  $\{X'_k\}$  respectively.*

*If there are no hyperlinks from  $W$  to  $V$  (i.e.  $H \cap (W \times V) = \emptyset$ ), then:*

- $\{X'_k\}$  is a  $(G', \omega', d, s')$  Random Search-Click process, where:

$$G' = (P', H') \tag{4.2}$$

$$P' = V \cup \{W\} \tag{4.3}$$

$$H' = \{(\rho(a), \rho(b)) : (a, b) \in H\} \tag{4.4}$$

$$\rho(a) = \begin{cases} a & \text{if } a \in V \\ W & \text{if } a = W \end{cases} \tag{4.5}$$

$$\omega'(a, b) = \begin{cases} \omega(a, b) & \text{if } a, b \in V \\ \sum_{w \in W} \omega(a, w) & \text{if } a \in V \text{ and } b = W \\ 0 & \text{if } a = W \text{ and } b \in V \\ 1 & \text{if } a = b = W \end{cases} \tag{4.6}$$

$$s'(a) = \begin{cases} s(a) & \text{if } a \in V \\ \sum_{w \in W} s(w) & \text{if } a = W \end{cases} \tag{4.7}$$

- $r(v) = r'(v)$  for all  $v \in V$ .
- $\sum_{w \in W} r(w) = r'(W)$ .

*Proof.* To prove this, I first show that  $\{X'_k\}$  is a Markov process, and then show that its transition function matches the one given above. The other consequences relating the PageRanks in the two processes follow trivially from that fact that a random variable can only take on one value (simultaneously).

I need some notation. Let  $\psi$  be the transition function of  $\{X_k\}$  and  $\psi'$  the transition function of  $\{X'_k\}$ .

I will now show that  $\{X'_k\}$  is a Markov process. Since there are no edges from  $W$  to  $V$  in  $G$ , we have  $\mathbf{P}(X_{k+1} = b | X_k = a) = ds(b) + 0 = ds(b)$  for all  $a \in W$  when  $b \in V$ . Similarly  $\mathbf{P}(X_{k+1} \in W | X_k = a) = 1 - d \sum_{v \in V} s(v)$  for all  $a \in W$ . Therefore, the conditional probability distributions for each state in  $W$  are the same and  $\{X'_k\}$  is a Markov process by *Theorem 3.11 (Lumpability Condition)*.

To show that,  $\{X'_k\}$  is a  $(G', \omega', d, s')$  Random Search-Click process, I need to show that the initial distribution  $s'$  and transition function  $\psi'$  derived from  $(G', \omega', d, s')$  match those of  $\{X'_k\}$ :

$$\mathbf{P}(\rho(X_0) = a) = s'(a) \tag{4.8}$$

$$\mathbf{P}(\rho(X_{k+1}) = b | \rho(X_k) = a) = \psi'(a, b) \tag{4.9}$$

The initial distribution matches trivially.

There are four cases to check for verifying the transition function matches. Firstly, when  $a, b \in V$ :

$$\mathbf{P}(\rho(X_{k+1}) = b \mid \rho(X_k) = a) = \mathbf{P}(X_{k+1} = b \mid X_k = a) \quad (4.10)$$

$$= \psi(a, b) \quad (4.11)$$

$$= ds(b) + (1 - d)\omega(a, b) \quad (4.12)$$

$$= ds'(b) + (1 - d)\omega'(a, b) \quad (4.13)$$

$$= \psi'(a, b) \quad (4.14)$$

Secondly, when  $a \in V$  and  $b = W$ :

$$\mathbf{P}(\rho(X_{k+1}) = W \mid \rho(X_k) = a) = \mathbf{P}(X_{k+1} \in W \mid X_k = a) \quad (4.15)$$

$$= \sum_{w \in W} \mathbf{P}(X_{k+1} = w \mid X_k = a) \quad (4.16)$$

$$= \sum_{w \in W} \psi(a, w) \quad (4.17)$$

$$= \sum_{w \in W} [ds(w) + (1 - d)\omega(a, w)] \quad (4.18)$$

$$= ds'(W) + (1 - d)\omega'(a, W) \quad (4.19)$$

$$= \psi'(a, W) \quad (4.20)$$

Thirdly, when  $a = W$  and  $b \in V$ :

$$\mathbf{P}(\rho(X_{k+1}) = b \mid \rho(X_k) = W) = \mathbf{P}(X_{k+1} = b \mid X_k \in W) \quad (4.21)$$

$$= ds(b) + (1 - d) \cdot 0 \quad (\text{no hyperlinks from inside } W \text{ to } b) \quad (4.22)$$

$$= ds'(b) + (1 - d)\omega'(W, b) \quad (4.23)$$

$$= \psi'(W, b) \quad (4.24)$$

Finally, when  $a = b = W$ :

$$\mathbf{P}(\rho(X_{k+1}) = W \mid \rho(X_k) = W) = \mathbf{P}(X_{k+1} \in W \mid X_k \in W) \quad (4.25)$$

$$= 1 - \mathbf{P}(X_{k+1} \in V \mid X_k \in W) \quad (4.26)$$

$$= 1 - \sum_{v \in V} \mathbf{P}(X_{k+1} = v \mid X_k \in W) \quad (4.27)$$

$$= 1 - \sum_{v \in V} [ds(v) + (1 - d) \cdot 0] \quad (4.28)$$

$$= 1 - d \sum_{v \in V} s(v) \quad (4.29)$$

$$= 1 - d(1 - \sum_{w \in W} s(w)) \quad (4.30)$$

$$= 1 - d(1 - s'(W)) \quad (4.31)$$

$$= ds'(W) + (1 - d) \cdot 1 \quad (4.32)$$

$$= \psi'(W, W) \quad (4.33)$$

So, I have now shown that  $\{X'\}$  is a  $(G', \omega', d, s')$  Random Search-Click process. I now need to prove that the above relationships between the PageRank functions  $r$  and  $r'$  hold.

Firstly, for  $v \in V$ :

$$r(v) = \lim_{k \rightarrow \infty} \mathbf{P}(X_k = v) \quad (4.34)$$

$$= \lim_{k \rightarrow \infty} \mathbf{P}(\rho(X_k) = v) \quad (4.35)$$

$$= \lim_{k \rightarrow \infty} \mathbf{P}(X'_k = v) \quad (4.36)$$

$$= r'(v) \quad (4.37)$$

Secondly, for  $W$ :

$$\sum_{w \in W} r(w) = \sum_{w \in W} \lim_{k \rightarrow \infty} \mathbf{P}(X_k = w) \quad (4.38)$$

$$= \lim_{k \rightarrow \infty} \mathbf{P}(X_k \in W) \quad (4.39)$$

$$= \lim_{k \rightarrow \infty} \mathbf{P}(\rho(X_k) = W) \quad (4.40)$$

$$= \lim_{k \rightarrow \infty} \mathbf{P}(X'_k = W) \quad (4.41)$$

$$= r'(W) \quad (4.42)$$

□

This following Lemma says that if two Random Search-Click processes are exactly the same outside of some subset  $V$  of all web pages, then you can shrink all other web pages down into one, and you get the same Random Search-Click processes out.

**Lemma 4.2 (PageRank Lumping Isomorphism).** *Let  $G_X = (V \cup W_X, H_X)$  and  $G_Y = (V \cup W_Y, H_Y)$  be webgraphs where  $V \cap W_X = V \cap W_Y = \emptyset$ . Let  $\{X_k\}$  be a  $(G_X, \omega_X, d, s_X)$  Random Search-Click process. Let  $\{Y_k\}$  be a  $(G_Y, \omega_Y, d, s_Y)$  Random Search-Click process.*

*Then  $\{X_k\}/W_X \cong \{Y_k\}/W_Y$  if the following conditions are satisfied:*

- *there are no hyperlinks from  $W_X$  to  $V$  nor from  $W_Y$  to  $V$ . (that is  $H_X \cap (W_X \times V) = H_Y \cap (W_Y \times V) = \emptyset$ )*
- *$\omega_X(a, b) = \omega_Y(a, b)$  for all  $a, b \in V$ .*
- *$s_X(a) = s_Y(a)$  for all  $a \in V$ .*

*Proof.* To prove that  $\{X_k\}/W_X \cong \{Y_k\}/W_Y$ , I construct an isomorphism between the resulting Random Search-Click processes. The requirement for no hyperlinks from pages in  $W_X$  and  $W_Y$  to pages in  $V$  is needed to use *Lemma 4.1 (PageRank Lumping)* to construct these Random Search-Click processes.

Let  $\{X'_k\} = \{X_k\}/W_X$ . By *Lemma 4.1 (PageRank Lumping)*,  $\{X'_k\}$  is a  $(G_{X'}, \omega_{X'}, d, s_{X'})$  Random Search-Click process (since there are no hyperlinks from pages in  $W_X$  to pages in  $V$ ), with:

$$G_{X'} = (P_{X'}, H_{X'}) \quad (4.43)$$

$$P_{X'} = V \cup \{W_X\} \quad (4.44)$$

$$\omega_{X'}(a, b) = \begin{cases} \omega_X(a, b) & \text{if } a, b \in V \\ \sum_{w \in W_X} \omega_X(a, w) & \text{if } a \in V \text{ and } b = W_X \\ 0 & \text{if } a = W_X \text{ and } b \in V \\ 1 & \text{if } a = b = W_X \end{cases} \quad (4.45)$$

$$s_{X'}(a) = \begin{cases} s_X(a) & \text{if } a \in V \\ \sum_{w \in W_X} s_X(w) & \text{if } a = W_X \end{cases} \quad (4.46)$$

Likewise,  $\{Y'_k\} = \{Y_k\}/W_Y$  is a  $(G_{Y'}, \omega_{Y'}, d, s_{Y'})$  Random Search-Click process:

$$G_{Y'} = (P_{Y'}, H_{Y'}) \quad (4.47)$$

$$P_{Y'} = V \cup \{W_Y\} \quad (4.48)$$

$$\omega_{Y'}(a, b) = \begin{cases} \omega_Y(a, b) & \text{if } a, b \in V \\ \sum_{w \in W_X} \omega_X(a, w) & \text{if } a \in V \text{ and } b = W_Y \\ 0 & \text{if } a = W_Y \text{ and } b \in V \\ 1 & \text{if } a = b = W_Y \end{cases} \quad (4.49)$$

$$s_{Y'}(a) = \begin{cases} s_Y(a) & \text{if } a \in V \\ \sum_{w \in W_Y} s_Y(w) & \text{if } a = W_Y \end{cases} \quad (4.50)$$

Now, consider the bijective function  $\xi : P_{X'} \rightarrow P_{Y'}$ :

$$\xi(a) = \begin{cases} a & \text{if } a \in V \\ W_Y & \text{if } a = W_X \end{cases} \quad (4.51)$$

To show that this is an isomorphism, I use the condition from *Theorem 3.8 (Markov Isomorphism Condition)*. This means it is sufficient to show that  $s_{X'}(a) = s_{Y'}(\xi(a))$  for all  $a \in P_{X'}$  and that  $\omega_{X'}(a, b) = \omega_{Y'}(\xi(a), \xi(b))$  for all  $a, b \in P_{X'}$ .

Firstly, it is clear from the assumptions that  $s_{X'}(a) = s_{Y'}(\xi(a))$  for all  $a \in V$ . For  $s_{X'}(W_X)$ , we have:

$$s_{X'}(W_X) = \sum_{w \in W_X} s_X(W_X) \quad (4.52)$$

$$= 1 - \sum_{w \in V} s_X(w) \quad (4.53)$$

$$= 1 - \sum_{w \in V} s_Y(w) \quad (4.54)$$

$$= s_{Y'}(W_Y) \quad (4.55)$$

$$= s_{Y'}(\xi(W_X)) \quad (4.56)$$

Secondly,  $\omega_{X'}(a, b) = \omega_{Y'}(\xi(a), \xi(b))$  for all  $a, b \in V$  by the assumption that  $\omega_X(a, b) = \omega_Y(a, b)$  on  $a, b \in V$ . The cases where  $a = W_X$  are trivial. The remaining case, where  $a \in V$  and  $b = W_X$  follows:

$$\omega_{X'}(a, W_X) = \sum_{w \in W_X} \omega_X(a, w) \quad (4.57)$$

$$= 1 - \sum_{v \in V} \omega_X(a, v) \quad (4.58)$$

$$= 1 - \sum_{v \in V} \omega_Y(a, v) \quad (4.59)$$

$$= \sum_{w \in W_Y} \omega_Y(a, W_Y) \quad (4.60)$$

$$= \omega_{Y'}(a, \xi(W_X)) \quad (4.61)$$

Therefore,  $\xi$  is an isomorphism, and  $\{X_k\}/W_X \cong \{Y_k\}/W_Y$ .  $\square$

This Lemma tells us the cost of attack of PageRank in the special case that the attacker does not include any hyperlinks to web pages not under his control. More precisely, it says that the sum of the attacker's webpages' PageRanks is exactly the probability of the random search function directing a random surfer to one of the attacker's web pages.

**Lemma 4.3 (PageRank Lump Sum).** Let  $G = (V \cup W, H)$  where  $V \cap W = \emptyset$ . Let  $\{X_k\}$  be a  $(G, \omega, d, s)$  Random Search-Click process. Let  $r$  be the PageRank function using the process  $\{X_k\}$ .

If there are no hyperlinks between  $V$  and  $W$  (i.e.  $H \cap (W \times V) = H \cap (V \times W) = \emptyset$ ), then:

- $\{Z_k\} = \{X_k\}/W/V$  is a  $(G_Z, \omega_Z, d, s_Z)$  Random Search-Click process, where:

$$G_Z = (P_Z = \{V, W\}, H_Z = \{(V, V), (W, W)\}) \quad (4.62)$$

$$\omega_Z(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \quad (4.63)$$

$$s_Z(A) = \sum_{a \in A} s(a) \quad (4.64)$$

- $\sum_{v \in V} r(v) = \sum_{v \in V} s(v)$  and  $\sum_{w \in W} r(w) = \sum_{w \in W} s(w)$ .

*Proof.* The first part of the proof is establishing that  $\{Z_k\}$  is the above Random Search-Click process. This follows from two applications of *Lemma 4.1 (PageRank Lumping)*. Once I know the transition function of  $\{Z_k\}$ , it is straight-forward to compute the PageRanks of a two-state Markov process.

Let  $\{Y_k\} = \{X_k\}/W$ . By *Lemma 4.1 (PageRank Lumping)*,  $\{Y_k\}$  is a  $(G_Y, \omega_Y, d, s_Y)$  Random Search-Click process, since there are no hyperlinks from pages in  $W$  to pages in  $V$ , where:

$$G_Y = (V \cup \{W\}, H \cap (V \times V) \cup (W, W)) \quad (4.65)$$

$$\omega_Y(a, b) = \begin{cases} \omega(a, b) & \text{if } a, b \in V \\ 0 & \text{if } a \in V \text{ and } b = W \\ 0 & \text{if } a = W \text{ and } b \in V \\ 1 & \text{if } a = b = W \end{cases} \quad (4.66)$$

$$s_Y = \begin{cases} s(a) & \text{if } a \in V \\ \sum_{w \in W} s(w) & \text{if } a = W \end{cases} \quad (4.67)$$

Moreover, the PageRanks are related by:

$$\sum_{v \in V} r(v) = \sum_{v \in V} r_Y(v) \quad (4.68)$$

Clearly,  $\{Z_k\} = \{Y_k\}/V = \{X_k\}/W/V$ . Since there are no hyperlinks from  $W$  to page in  $V$ , by *Lemma 4.1 (PageRank Lumping)*,  $\{Z_k\}$  is a  $(G_Z, \omega_Z, d, s_Z)$  Random Search-Click process (where  $G_Z, \omega_Z$  and  $s_Z$  are defined above). Moreover, the PageRanks are related by:

$$\sum_{v \in V} r_Y(v) = r_Z(V) \quad (4.69)$$

By Definition (Random Search-Click Process), the transition function  $\psi_Z$  of  $\{Z_k\}$  is:

$$\psi_Z(a, b) = ds_Z(b) + (1 - d)\omega_Z(a, b) \quad (4.70)$$

$$= \begin{cases} 1 - ds_Z(W) & \text{if } a = V \text{ and } b = V \\ ds_Z(W) & \text{if } a = V \text{ and } b = W \\ ds_Z(V) & \text{if } a = W \text{ and } b = V \\ 1 - ds_Z(V) & \text{if } a = W \text{ and } b = W \end{cases} \quad (4.71)$$

Now, by *Corollary 3.6 (Two-state Convergence)*, I can compute the PageRank of  $V$  in  $G_Z$ :

$$\sum_{v \in V} r(v) = \sum_{v \in V} r_Y(v) \quad (4.72)$$

$$= r_Z(V) \quad (4.73)$$

$$= \lim_{k \rightarrow \infty} \mathbf{P}(Z_k = V) \quad (4.74)$$

$$= \frac{\psi_Z(W, V)}{\psi_Z(W, V) + \psi_Z(V, W)} \quad (4.75)$$

$$= \frac{ds_Z(V)}{ds_Z(V) + ds_Z(W)} \quad (4.76)$$

$$= \frac{ds_Z(V)}{d} \quad (4.77)$$

$$= s_Z(V) \quad (4.78)$$

$$= \sum_{v \in V} s(v) \quad (4.79)$$

The same relationship holds for  $W$  trivially:

$$\sum_{w \in W} r(W) = 1 - \sum_{v \in V} r(v) \quad (4.80)$$

$$= 1 - \sum_{v \in V} s(v) \quad (4.81)$$

$$= \sum_{w \in W} s(w) \quad (4.82)$$

□

Finally, I am up to the Cost of Attack Theorem. Suppose an attacker has a set of web pages  $V$  and the rest of the internet is made up of a set of web pages  $W$ . If there are no hyperlinks from pages in  $W$  to pages the attacker controls,  $V$ , then the Cost of Attack Theorem says that the sum of the attacker's pages' PageRanks is less than or equal to the proportion of initial votes allocated to the attackers' web pages.

**Theorem 4.4 (PageRank Cost of Attack).** *Let  $G = (V \cup W, H)$  be a webgraph where  $V \cap W = \emptyset$ . Let  $\{X_k\}$  be a  $(G, \omega, d, s)$  Random Search-Click process with transition function  $\psi$ . Let  $r$  be the PageRank function using the process  $\{X_k\}$ .*

*If there are no hyperlinks from  $W$  to  $V$  (i.e.  $H \cap (W \times V) = \emptyset$ ), then:*

$$\sum_{v \in V} r(v) \leq \sum_{v \in V} s(v) \quad (4.83)$$

*Moreover, equality occurs if there are no hyperlinks from  $V$  to  $W$ .*

*Proof.* To prove this, I reduce this general case to the special case already proven in *Lemma 4.3 (PageRank Lump Sum)*. I do this by using *Lemma 4.2 (PageRank Lumping Isomorphism)* to rearrange the pages in  $W$  into a convenient form that allows me to compute the amount of PageRank donated by pages in  $V$  to pages in  $W$ .

Here, I construct the convenient form that has a lumped process isomorphic to  $\{X_k\}/W$ .

Let  $\{Y_k\}$  be the  $(G_Y, \omega_Y, d, s_Y)$  Random Search-Click process with:

$$G_Y = (P_Y, H_Y) \quad (4.84)$$

$$P_Y = V \cup \{\mu, \nu\} \quad (4.85)$$

$$H_Y = \{(h, h)\} \cup \{(\rho_Y(a), \rho_Y(b)) : (a, b) \in H\} \quad (4.86)$$

$$\rho_Y(a) = \begin{cases} a & \text{if } a \in V \\ \mu & \text{if } a \in W \end{cases} \quad (4.87)$$

$$\omega_Y(a, b) = \begin{cases} \omega(a, b) & \text{if } a, b \in V \\ \sum_{w \in W} \omega(a, w) & \text{if } a \in V \text{ and } b = \mu \\ 0 & \text{if } a \in \{\mu, \nu\} \text{ and } a \neq b \\ 1 & \text{if } a = b = \mu \text{ or } a = b = \nu \end{cases} \quad (4.88)$$

$$s_Y(a) = \begin{cases} s(a) & \text{if } a \in V \\ 0 & \text{if } a = \mu \\ \sum_{w \in W} s(w) & \text{if } a = \nu \end{cases} \quad (4.89)$$

Let  $\{X'_k\} = \{X_k\}/W$  and  $\{Y'_k\} = \{Y_k\}/\{\mu, \nu\}$ .

Since in  $G$  and  $G_Y$ , there are no edges from  $W$  to  $V$  and  $\{\mu, \nu\}$  to  $V$  respectively, we have by *Lemma 4.1 (PageRank Lumping)*:

$$\sum_{v \in V} r(v) = \sum_{v \in V} r_{X'}(v) \quad (4.90)$$

$$\sum_{v \in V} r_Y(v) = \sum_{v \in V} r_{Y'}(v) \quad (4.91)$$

By *Lemma 4.2 (PageRank Lumping Isomorphism)*,  $\{X'_k\} \cong \{Y'_k\}$  since  $\omega(a, b) = \omega_Y(a, b)$  for all  $a, b \in V$ , and  $s(a) = s_Y(a)$  for all  $a \in V$ . Therefore, we have:

$$\sum_{v \in V} r_{X'}(v) = \sum_{v \in V} r_{Y'}(v) \quad (4.92)$$

Since there are no edges between  $V \cup \{\mu\}$  and  $\{\nu\}$ , *Lemma 4.3 (PageRank Lump Sum)* applies:

$$\sum_{v \in (V \cup \{\mu\})} r_Y(v) = \sum_{v \in (V \cup \{\mu\})} s_Y(v) \quad (4.93)$$

$$= \sum_{v \in V} s(v) \quad (4.94)$$

Finally, we can obtain the desired inequality:

$$\sum_{v \in V} r(v) = \sum_{v \in V} r_{X'}(v) \quad (4.95)$$

$$= \sum_{v \in V} r_Y(v) \quad (4.96)$$

$$= \sum_{v \in V \cup \{\mu\}} r_Y(v) - r_Y(\mu) \quad (4.97)$$

$$= \sum_{v \in V} s(v) - r_Y(\mu) \quad (4.98)$$

$$\leq \sum_{v \in V} s(v) \quad (4.99)$$

The condition for equality has already been proved in *Lemma 4.3 (PageRank Lump Sum)*. □

## 4.2 Cost of Attack of Google

In the previous section, I proved that PageRank’s initial vote allocation controls the cost of attack. In this section, I conjecture how Google parameterizes PageRank and use this to estimate the cost of attacking Google by buying domain names. I then compare this with empirical evidence of the cost of attack by trading hyperlinks.

### 4.2.1 Adapting the Theorem

In [Page et al., 1998], the Google founders suggest allocating initial votes to the root web pages on each server on the internet. These are usually called `/index.html`. For example `http://eff.org/index.html` is a root web page, but not `http://eff.org/archive.html`.

It is unclear how Google determines what a server is. One strategy might be to compare IP (Internet Protocol) addresses. Another might be to compare domain names. [Levien, 2003] suggests that domain names<sup>1</sup> are a good option because they cost money. The prices paid can probably be inferred from the internet’s `whois` records. These prices could be used to weight the initial votes.

For the purposes of estimating the cost of attack of Google, I assume that Google allocates the initial votes to root web pages on domains, weighting by cost paid. I also assume that the attacker plans to attack by buying domain names, creating web pages and adding hyperlinks to web pages he controls. In particular, I assume the attacker has no hyperlinks from web pages he does not control. Clearly, there are many other ways to attack such as hacking into the domain name registrars or hacking into high PageRank websites. These are beyond the scope of my estimates. However, these attacks are probably quite expensive.

This corollary applies the theory developed in the previous section to the domain name configuration of PageRank. It says that the sum of the attacker’s web pages’ PageRanks is less than or equal to the proportion of money the attacker spent on acquiring domain names over the total amount paid by everyone on domain names. This means that the size and the link structure of the attacker’s web pages are irrelevant. ‘Link farms’ are futile.

**Corollary 4.5 (Domain-name Cost of Attack).** *Let  $G = (P, H)$  be a webgraph where  $P = V \cup W$  and  $V \cap W = \emptyset$ . Let  $s(a) = \frac{c(a)}{\sum_{p \in P} c(p)}$ , where  $c(p)$  is the cost of registering page  $p$ . Let  $\{X_k\}$  be a  $(G, \omega, d, s)$  Random Search-Click process for some click distribution  $\omega$  and search probability  $d$ .*

*If there are no hyperlinks from pages in  $W$  to pages in  $V$ , then:*

$$\sum_{v \in V} r(v) \leq \frac{\sum_{v \in V} c(v)}{\sum_{p \in P} c(p)}$$

*Moreover, equality occurs if there are no hyperlinks from pages in  $V$  to pages in  $W$ .*

*Proof.* This is a trivial application of *Theorem 4.4 (PageRank Cost of Attack)*:

$$\begin{aligned} \sum_{v \in V} r(v) &\leq \sum_{v \in V} s(v) \\ &= \sum_{v \in V} \frac{c(v)}{\sum_{p \in P} c(p)} \\ &= \frac{\sum_{v \in V} c(v)}{\sum_{p \in P} c(p)} \end{aligned}$$

The theorem provides the same conditions for equality. □

---

<sup>1</sup>The Domain Name System is administered by the Internet Corporation for Assigned Names and Numbers. <http://www.icann.org/>

Item	Source	Lower Estimate	Upper Estimate
Number of paid domains	Netcraft	20000000	45000000
Cost of a domain	various providers	\$US10 / year	\$US20 / year
Money spent on domains	<b>number</b> $\times$ <b>cost</b>	\$US200M / year	\$US900M / year
UserPageRank log base	deduced from Google	10	10
Highest PageRank	[Pandurangan et al., 2002]	0.0001	0.001
$E = \frac{1}{d} - 1$	[Page et al., 1998]	0.15	0.15

Table 4.1: Estimates for the figures required to compute the cost of attack of Google. In all cases except UserPageRank log base, higher estimates lead to a higher estimated cost of attack.

This corollary means that I only need to consider the fraction

$$\frac{\sum_{v \in V} c(v)}{\sum_{p \in P} c(p)}$$

when estimating the cost of attack. To evaluate the cost of attack of a particular PageRank level, I need to compute:

$$\text{cost} = \text{PageRank} \cdot \sum_{p \in P} c(p)$$

## 4.2.2 Estimating the Cost of Attack

To estimate the cost of attack of a particular PageRank, the total cost of registering all domain names is needed. Moreover, Google presents PageRanks on a log scale. To be useful, the cost of attack needs to be presented in this form also. Unfortunately, the total cost of domain names and the base and constant of the log are not readily available, so I estimate these values. The estimates of the relevant numbers are summarized in *Table 4.1*.

To estimate the total cost of all domain names, I multiply an estimate of the number of domain names by an estimate of the cost of domain names. According to Netcraft,<sup>2</sup> there are roughly 45 000 000 domain names on the internet. However, many of these sites are unused placeholder domain names that Google possibly ignores. Netcraft also reports that 20 000 000 domains have different content. The number of domains that Google counts probably lies in between.

Top level domain names cost about \$US10 per year.<sup>3</sup> Many domain names cost much more, so this is probably a conservative estimate.

Google only displays PageRank scores on a log scale, with the highest ranking page normalized to 10. Google probably computes this with:

$$\text{LogPageRank}(a) = 10 + \log_b \frac{r(a)}{\max r}$$

For comparison purposes, it would be helpful to compute the cost of PageRanks on this scale. Unfortunately, neither the PageRank paper [Page et al., 1998] nor the Google website give any information about either of the the log base  $b$  or the maximum PageRank  $\max r$ . Therefore, I estimate them.

[Pandurangan et al., 2002] have computed the PageRanks on their own university’s web site with 100 000 pages, and on the WT10G corpus [Bailey et al., 2004] of 1.69 million web pages from the world wide web. The highest PageRank in the university graph was about 0.0005 while the highest PageRank in the WT10G graph was 0.001. These numbers are similar despite the order of magnitude difference in the corpus sizes. [Pandurangan et al., 2002] argues that the subsets of the World Wide Web share the same structure as the

<sup>2</sup><http://www.netcraft.com>

<sup>3</sup>For example, <http://www.active-domain.com/> offers domain names for \$US8.50 per year.

UserPageRank	Cost (\$US / year)	
	Lower Estimate	Upper Estimate
4	\$0.02	\$0.90
5	\$0.20	\$9
6	\$2	\$90
7	\$20	\$900
8	\$200	\$9000
9	\$2000	\$90000

Table 4.2: Estimated cost of attack of Google, where PageRanks are log-scale.

whole, and share similar PageRank distributions. Therefore, I can conservatively assume that the highest PageRank on the World Wide Web is between 0.0001 and 0.001.

To deduce the the log base  $b$ , I found a pair of pages with a simple link structure. I found a web page  $\mu^4$  with UserPageRank 2.5.  $\mu$  is only linked from  $\nu^5$ .  $\nu$  has a UserPageRank of 5 and 302 outgoing links.

Note that the UserPageRank numbers were read off a bar meter from the Google ToolBar web browser plugin. This bar meter is also available in HTML for pages listed on the Google Directory. In the HTML version, the bar grows in increments of 0.25 UserPageRank points. Therefore, a difference in UserPageRank of 2.5 should be significant enough to avoid large errors.

Assuming Google uses PageRank2 with  $E = 0.15$ , I deduce by by *Corollary 3.27 (PageRank2 is a Special Case)*:

$$r(\mu) = \frac{1}{0.15 + 1} \cdot \frac{r(\nu)}{302} \approx \frac{r(\nu)}{347}$$

By substitution, I derive the base  $b$ :

$$\begin{aligned} r(\mu) &= \frac{r(\nu)}{302} \\ 10 + \log_b \frac{r(\mu)}{\max r} &= 10 + \log_b \frac{r(\nu)}{302 \cdot \max r} \\ 10 + \log_b \frac{r(\mu)}{\max r} &= 10 + \log_b \frac{r(\nu)}{\max r} - \log_b 347 \\ \text{UserPageRank}(\mu) &= \text{UserPageRank}(\nu) - \log_b 347 \\ 2.5 &= 5 - \log_b 347 \\ b &= 10.382 \approx 10 \end{aligned}$$

I conclude that Google uses a base  $b = 10$  and has a maximum PageRank of between 0.0001 and 0.001. My estimated costs of attack from this information are in *Table 4.2*. Of particular interest, a company buying a single domain name for \$US10 will automatically achieve a UserPageRank of between 5 and 6.5.

### 4.2.3 Comparison with Empirical Cost of Attack

A market has developed for trading hyperlinks carrying PageRank from the seller to the buyer. Apparently, owners of some web pages value their PageRanks less than what others are prepared to pay. Information from this market might be helpful in comparing my estimated cost of attack with this market cost.

SearchKing<sup>6</sup> is Google PageRank broker, that matchmakes buyers and sellers of PageRank. Sellers add links to buyers' sites, increasing the buyers' sites' PageRanks.

<sup>4</sup><http://www.cs.mu.oz.au/~clai>. I verified this has only one incoming hyperlink with Google's link: search feature.

<sup>5</sup>[http://www.cs.mu.oz.au/~fjh/cs\\_homepages.html](http://www.cs.mu.oz.au/~fjh/cs_homepages.html)

<sup>6</sup><http://www.searchking.com>

Seller's UserPageRank	Buy (\$US / year)	Sell (\$US / year)
4	\$348	\$180
5	\$588	\$300
6	\$1188	\$600
7	\$4188	\$2100
8	\$9588	\$4800
9	\$CALL	\$CALL

Table 4.3: SearchKing's price list for Google PageRanks (log scale). This is misleading, because it lists the PageRank of the seller's page rather than the UserPageRank the buyer would acquire. The difference depends on how many outgoing links the seller's page has. Source: <http://www.pradnetwork.com/pricing.htm> (gives monthly pricing) and <http://www.pradnetwork.com/inventory.htm> (states that sellers get half), which is part of <http://www.searchking.com>

In the past, SearchKing has sought other strategies such as buying domain names of unmaintained web sites, and exploiting stale links to the old site to promote new sites. Google responded to this by removing ellapsed domain names from the set of initial votes. SearchKing thought this was unfair, and sued Google. The case was thrown out of court. <sup>7</sup>

SearchKing's price list appears in *Table 4.3*. The price list is radically different from my cost of attack estimates. SearchKing has an enormous markup of 50% between buying and selling.

The SearchKing price list seems to grow exponentially with base 3, while my own estimates grow exponentially with base 10. This means the market significantly overvalues low PageRanks and less significantly undervalues high PageRanks.

This discrepancy can not be explained by market factors such as differences in supply and demand. My estimates have the value of UserPageRank 8 as 1000 times as much as a UserPageRank of 5. Therefore, a seller should choose to sell 1000 hyperlinks contributing a UserPageRank of 5 to SearchKing at \$US 300 / year each, rather than 1 hyperlink at \$US 4800 / year.

One possible explanation is the market does not understand PageRank, and is irrational. Another possibility is that my estimate of 10 for the log base for PageRank is incorrect and should be closer to 3. Finally, high PageRank pages with thousands of outgoing links may be conspicuous to Google. Therefore, they may be more difficult to exploit and hence less valuable.

### 4.3 Limitations

While the Cost of Attack Theorem does give insight into the behaviour of PageRank, it is inadequate as a proof of attack resistance.

The assumption that the attacker receives no hyperlinks from the original web pages is a problem. As the theorem currently stands, when an attacker receives one such hyperlink, it can provide an upper bound on neither the sum of the attacker's PageRanks, nor on the cost of acquiring a higher PageRank. However, I believe this theory can be generalized, and this will not be a practical problem.

More fundamentally, there usually many attacks available to an attacker. For example, an attacker could break into a computer containing a page with high PageRank and add some hyperlinks to the attacker's pages. However, the cost of these attacks depend on the reputation system rather than PageRank. The Cost of Attack Theorem does capture all of the attacks that cost the attacker nothing: adding and modifying web pages that the attacker owns.

---

<sup>7</sup>Stefanie Olsen, "Judge Dismisses Suit Against Google", *CNET News*, 30 May 2003, <http://news.com.com/2100-1032-3-1011740.html>

## 4.4 Implications

If the upper estimates in *Table 4.2.2* were accurate, then perhaps Google could report that a web page with a PageRank of 9 has a reputation value of \$US 90 000 per year. Perhaps this means that the natural unit of PageRank on Google is not a probability between 0 and 1, nor a log scaled number between 0 and 10, but rather dollars per year.

Perhaps a consumer deciding whether to trust an online book vendor could conclude that a reputation level costing \$US 90 000 per year would not be worth buying merely to rip people off. A scam of that proportion would be very risky and has a high up-front cost barrier. Therefore, the vendor most likely achieved a high reputation through genuine recommendations.

While this possibility is tantalizing, it is distant. Firstly, only the web pages with particularly high PageRanks are likely to have a high enough cost of attack for consumers to dismiss attack as a possibility. This might mean that PageRank is not telling consumers anything about reputation that they did not already know.

Secondly, PageRank's lack of complaints means the payoff from an attack is bounded only by the time for which the acquired domain names last. If there is no mechanism for revoking PageRank, then perhaps \$US 90 000 is a small price to pay to easily scam thousands of people.

Finally, acquisition of high Google PageRanks at present is purely motivated by attracting web surfers. Few web surfers pay any attention to PageRank to assist in evaluating trustworthiness. If PageRank became important for consumer decisions, then perhaps no-one would ever spend any of their PageRank on recommending others.

## 4.5 Future Work

There are several areas that are worthy of further research.

Most obviously, relaxing the limitations of the cost of attack theorem would provide greater confidence in the cost of attack measurements or predictions.

Secondly, if the cost of attack of PageRank is to ever be equated with the value of reputation, then complaints are essential. Therefore, research into adding complaints in a reasonable way to PageRank would be useful. As I discussed in *Section 2.5*, this is a difficult problem.

Finally, if these first two problems were solved, then a Game Theoretic analysis of communities using PageRank to make decisions could provide a sophisticated understanding of the strengths and weaknesses of PageRank. In particular, it might describe which circumstances PageRank can and can not be trusted, and give practical advice to consumers.

# Chapter 5

## Contribution

My original contributions in this thesis include:

- I clarified the definition of *reputation* and *online reputation system*, and introduced the term *reputation metric*.
- My definition and presentation of PageRank has some novel aspects:
  - I suggest that pages with no outgoing hyperlinks should include a self-hyperlink for the purposes of PageRank. This is the natural ‘selfish’ thing to do.
  - I suggest the terminology *random search function*. Earlier work only described this in terms of a random jump when a surfer gets bored.
- I noticed and proved that the original definition of PageRank in [Brin and Page, 1998] to be inconsistent with the definition in [Page et al., 1998]. However, correct definitions of PageRank are also widespread... it appears these authors assumed the original definition is equivalent.
- I proved that the correct definition of PageRank is well defined.
- I presented a formal proof that algebraic definitions of PageRank match the Random Surfer model definition. Previously, only rough eigenvector arguments were used (as in [Page et al., 1998]), which are sound, but leave out lots of details.
- I proposed and proved *Theorem 4.4 (PageRank Cost of Attack)*, and the three preceding lemmas: *Lemma 4.1 (PageRank Lumping)*, *Lemma 4.2 (PageRank Lumping Isomorphism)* and *Lemma 4.3 (PageRank Lump Sum)*. [Levien, 2003] first conjectured that a theorem like this might exist.
- I showed how my theory can be applied to estimate the cost of attack of reputation systems, including Google.

I independently re-discovered the standard results *Corollary 3.6 (Two-State Equilibrium)* and a special case of *Theorem 3.11 (Lumpable Condition)*, neither of which appeared in the text book (namely, [Behrends, 1999]) I used to learn Markov theory. I had to consult several mathematicians to find the magic word *lumpability*. Prior to this, I proved *Lemma 4.3 (PageRank Disjoint Lump Sum)* inspired by the Deterministic Finite Automata minimization algorithm.

Neither of my supervisors have a background in Markov theory, PageRank nor reputation systems.

## Chapter 6

# Conclusions

When strangers meet on the internet, reputation is most likely the only feasible way they can establish trust. For a reputation system to be useful for establishing trust, it must have a high cost of attack.

Google's PageRank algorithm is promising because it weights all recommendations by the reputation of the sources. I have examined PageRank and proved that under some limited conditions, PageRank has a cost of attack equal to the cost of acquiring PageRank's initial votes. However, this high cost of attack is compromised by PageRank's lack of support for complaints. Attackers can profit highly from buying a high PageRank because deception goes unpunished.

An interesting implication of understanding cost of attack is that reputation levels can be assigned market values. My estimates of the cost of attack of Google are radically different from market valuation. The market appears to significantly overvalue low PageRanks and marginally undervalue high PageRanks.

Markets can remain irrational longer than you can remain solvent.

— John Maynard Keynes

## Acknowledgements

Paul Gruba, Liz Sonenberg, Graham Byrnes, Catherine Lai, Peter Taylor, Sanming Zhou, Suelette Dreyfus, Peter Eckersley, Geordie Zhang and Sam Joseph spent a lot of time providing me with useful feedback.

# Bibliography

- [Aberer and Despotovic, 2001] Aberer, K. and Despotovic, Z. (2001). Managing Trust in a Peer-2-Peer Information System. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*.
- [Bailey et al., 2004] Bailey, P., Craswell, N., and Hawking, D. (2004). Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*.
- [Behrends, 1999] Behrends, E. (1999). *Introduction to Markov Chains (with Special Emphasis on Rapid Mixing)*. Vieweg Verlag.
- [Bianchini et al., 2004] Bianchini, M., Gori, M., and Scarselli, F. (2004). Inside PageRank. *ACM Transactions on Internet Technology*.
- [Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.
- [Calkins, 2001] Calkins, M. (2001). My reputation always had more fun than me: The failure of eBay’s feedback model to effectively prevent online auction fraud. *Richmond Journal of Law and Technology*, 7.
- [Douceur, 2002] Douceur, J. (2002). The Sybil attack. In *1st International Workshop on Peer-to-Peer Systems (IPTPS’02)*.
- [Haveliwala, 2002] Haveliwala, T. (2002). Topic-sensitive pagerank. In *The Eleventh International World Wide Web Conference*.
- [Kamvar et al., 2003] Kamvar, S., Schlosser, M., and Garcia-Molina, H. (2003). The EigenTrust algorithm for reputation management in P2P networks. In *The Twelfth International World Wide Web Conference*.
- [Kemeny and Snell, 1976] Kemeny, J. G. and Snell, J. L. (1976). *Finite Markov Chains*. Springer-Verlag.
- [Levien, 2003] Levien, R. (2003). Attack resistant trust metrics. Draft PhD Thesis, University of California at Berkeley.
- [Marsh, 1994] Marsh, S. (1994). *Formalising Trust as a Computational Concept*. PhD thesis.
- [Mui et al., 2001] Mui, L., Halberstadt, A., and Mohtashemi, M. (2001). A computation model of trust and reputation. In *Proceedings of the 35th Hawaii International Conference on System Sciences*.
- [Mui et al., 2002] Mui, L., Halberstadt, A., and Mohtashemi, M. (2002). Notions of reputation in multi-agent systems: A review. In *Proceedings of the First Joint International Conference on Autonomous Agents and Multi-Agent Systems*.
- [Norris, 1997] Norris, J. (1997). *Markov Chains*. Cambridge University Press.
- [Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

- [Pandurangan et al., 2002] Pandurangan, G., Raghavan, P., and Upfal, E. (2002). Using PageRank to Characterize Web Structure. In *8th Annual International Computing and Combinatorics Conference (COCON)*.
- [Resnick et al., 1994] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., and Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina. ACM.
- [Resnick and Zeckhauser, 2002] Resnick, P. and Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of eBay’s reputation system. the economics of the internet and e-commerce. *Advances in Applied Microeconomics*, 11.
- [Resnick et al., 2000] Resnick, P., Zeckhauser, R., Friedman, E., and Kuwabara, K. (2000). Reputation systems. *Communications of the ACM*, 43(12):45 – 58.
- [Sabater and Sierra, 2002] Sabater, J. and Sierra, C. (2002). Reputation and social network analysis in multi-agent systems. In *Proceedings of the First Joint International Conference on Autonomous Agents and Multi-Agent Systems*.
- [Yolum and Singh, 2003] Yolum, P. and Singh, M. (2003). Emergent properties of referral systems. In *Second International Conference on Autonomous Agents and Multiagent Systems (AAMAS03)*.
- [Zacharia et al., 1999] Zacharia, G., Moukas, A., and Maes, P. (1999). Collaborative reputation mechanisms in electronic marketplaces. In *Proceedings of the 32nd Hawaii International Conference on Systems Sciences*.