Robust Reputations for Peer-to-peer Markets

Jonathan David Traupman

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-75
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-75.html

May 24, 2007

Robust Reputations for Peer-to-peer Markets

by

Jonathan David Traupman

B.S. Yale University 1996
M.S. University of California, Berkeley 2002

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor J. D. Tygar, Chair
Professor John Canny
Professor Hal Varian

Spring 2007

The dissertation of Jonathan David Traupman is approved.

Chair	Date
	Date
	Date

University of California, Berkeley

Spring 2007

Robust Reputations for Peer-to-peer Markets

by

Jonathan David Traupman

Abstract

Robust Reputations for Peer-to-peer Markets

by

Jonathan David Traupman

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor J. D. Tygar, Chair

This thesis investigates on-line reputation systems for peer-to-peer markets and presents a number of systems which increase robustness against attacks on individual reputations while still distinguishing between honest and dishonest user behavior.

The fluidity of identity on-line and the unavailability of practical legal recourse make evaluating trust and risk in on-line markets both vital and difficult. Reputation systems have been proposed as one possible means of building trust among strangers by aggregating the experience of many users, and prove more-or-less effective in peer-to-peer marketplaces like eBay. However, the very attributes that make reputation systems helpful also make them a target for fraud. A good reputation is valuable, so some users may try to circumvent the system to gain a high reputation without effort. We look at two specific ways in which users attack reputation systems in peer-to-peer markets and discuss ways in which the damage can be mitigated.

We first address retaliatory negative feedback, where a user leaves a negative feedback for someone who complained about their behavior. We show that allowing retaliation can result in a reputation system that is incapable of identifying low-quality users and allows cheating to go unpunished. We then present EM-Trust, a system that is better able to estimate true user quality even with high levels of retaliation.

We next look at the issue of sybil attacks, where a single user creates a large collection of identities to increase his own reputation. We show that EigenTrust, a widely discussed algorithm that purports to resist similar collusion attacks, does not work against sybils. We then present Relative Rank, a transformation of EigenTrust that is both sybil resistant and better suited to peer-to-peer marketplaces. Finally, we discuss RAW, a variation of PageRank that offers additional guarantees of sybil-resistance.

We demonstrate that it is possible to design reputation systems that are as effective as existing non-robust ones at discriminating between honest and dishonest user behavior, and considerably less affected by common attacks against these systems.

Professor J. D. Tygar
Dissertation Committee Chair

For my wife, Mayumi

and

For Mom and Dad

Contents

Acknowledgements

Most of the research that went into this thesis was performed under the guidance of Professor Robert Wilensky, my research advisor for the majority of my time at Berkeley. Due to an unfortunate illness, Robert was unable to serve on the committee for this thesis, but his many contributions are present throughout this work. I am deeply grateful for all of his support and guidance over the past several years.

I would also like to extend my gratitude to Professor Doug Tygar, who, without being asked, voluteered to chair my thesis committee and to provide financial support for my final semester at Berkeley. His generosity helped to avert what could have been a graduate student's worst nightmare.

Many thanks as well to the other members of my committee, Professor John Canny and Professor Hal Varian. Their advice, suggestions, and criticisms helped to shape and strengthen this work considerably.

The best part of my graduate program was the opportunity to meet and work with the tremendous number of extremely intelligent, talented, and simply fun people here at Berkeley. So, a big "thank you" to all of the faculty, fellow students, and visitors with whom I studied, collaborated, and socialized over the past eight years.

Finally, thank you to my family — my wife, Mayumi, my parents, Arnold and Barbara, and my siblings Matt, Gabe, and Emily — and to my many friends for their support, encouragement, and advice and for not asking "so just how long are you going to be in school, anyway?" more often than necessary.

This work was supported in part by TRUST (The Team for Research in Ubiquitous Secure Technology), which receives support from the National Science Foundation (NSF award number CCF-0424422) and the following organizations: AFOSR (#FA9550-06-1-0244) Cisco, British Telecom, ESCHER, HP, IBM, iCAST, Intel, Microsoft, ORNL, Pirelli, Qualcomm, Sun, Symantec, Telecom Italia and United Technologies. Additional research in this thesis was supported by the Digital Libraries Initiative under grant NSF CA98-17353. Computing resources were provided by the Petabyte Storage Infrastructure, which received support from the National Science Foundation

(award number 0303575). The opinions in this dissertation are those of the author and do not necessarily reflect those of the US government, any of its agencies, or any funding sponsor.

Earlier versions of portions of the work in this thesis have been published previously in (105), (107), and (106).

Chapter 1

Introduction

This thesis investigates on-line reputation systems for peer-to-peer markets and presents a number of systems which increase robustness against attacks on individual reputations while still distinguishing between honest and dishonest user behavior.

The success of most human interactions is predicated on some notion of trust. Nearly any transaction between two people or groups, be it financial or social, entails some risk that one side or the other will not fulfill their end of the bargain. The trust each side has in the other is a key factor in estimating the risk of failure and thus the interaction's expected worth. The problem of trust is a fundamental one, which our society has addressed by building durable legal, financial, and social edifices whose main goals are to foster trust between individuals, to mitigate the damage done when a trusted party fails, and to punish those who deliberately try to usurp these systems at the expense of others. Most of the time, we are not even conscious of the ways in which society works to facilitate trust between its members.

The rise of effortless global communications — most notably, the Internet — has done nothing to reduce the ancient need to trust those which whom we choose to interact. However, it has strained traditional methods' ability to create and maintain trust. Most existing systems make some assumptions about locality or identity that simply do not hold in the largely anonymous, globally distributed "community" of the Internet. For example, an untrustworthy individual can easily erase a bad reputation by creating a new on-line identity. (45) These cheap pseudonyms also allow the

creation of entirely faked transaction histories, further complicating the problem of distinguishing trustworthy and untrustworthy individuals. Even with a persistent identity, the essentially global nature of the Internet permits services to locate themselves in favorable regulatory regimes so as to skirt inconvenient local laws. (47)

Even if possible, trying to regulate this freedom and fluidity risks doing more harm than good. The ease with which information and services can be created and globally disseminated is a key component of the economic engine driving growth on-line. Even the most troublesome technologies typically have positive uses: the same techniques that shield the identities of criminals conducting on-line auction fraud or trading child pornography also permit dissidents to avoid censorship by repressive regimes and individuals to discuss personal issues anonymously. While the Internet clearly disrupts many of the trust-building methods and institutions that have worked well in the past, we believe that trying to close Pandora's box is not the solution. Rather, we need new mechanisms for building and maintaining trust on-line, and these mechanisms need to be robust against the type of use and abuse they will receive in this environment.

We concentrate on one such mechanism, the reputation system. The remainder of this chapter introduces reputation systems and one particular domain that has adopted them widely: on-line peer-to-peer markets, sites where individual buyers and sellers come to trade without intermediaries. We also summarize the methods by which users try to circumvent these systems in order to become more trusted than they otherwise should be. Subsequent chapters delve into specific details of several of these attacks and propose algorithms that effectively resist them. It is our belief that robust reputation systems can be as effective at building trust on-line as traditional methods are in the off-line world. The importance of such systems will only increase as both commerce and socialization becomes more and more decentralized and global.

1.1 Introduction to Reputation Systems

All reputation systems, regardless of their approach and application domain, aim to solve one major problem: how to reliably assist users in making trust decisions about otherwise unknown strangers. The ideal reputation system would be an oracle that predicts with 100% precision and

accuracy whether or not a potential partner to a transaction will behave honestly[1] or not. Of course, such a perfect system is decades in the future, if not outright impossible. However, it is helpful to keep this goal in mind as we design real world systems.

Before examining reputation system mechanisms and applications in more detail, it is helpful to spend a short while looking at how and why people make trust decisions. The need for trust arises when an individual must make a decision under uncertainty that entails some risk. For example, the decision to make one's first solo skydiving jump is obviously risky. There is also considerable uncertainty: will the parachute open? Is the plane flying high enough? Are there any crosswinds that will carry me into a power line? One can reduce the uncertainty through proper procedures, but ultimately one's decision to jump will require trust in the parachute's manufacturer, the person that packed it, the pilot, the skills of the instructor, and many other people and things. Unless one lives in a bubble, we have to make similar trust decisions many times daily, though usually the risks are not quite so extreme.

Of particular interest in this thesis are decisions involving direct interaction with another individual. Such interactions are common in commerce, collaborative work, social interactions, or any other endeavor where people, companies, or machines join together for some common or mutually beneficial goal. Because the intentions of the other party are essentially unknowable, both sides[2] must decide whether they wish to interact based on their trust in each other.

There are several approaches to making this trust decision, based on the relationship the individuals have. The most straightforward mechanism is to rely on personal experience. This is a method we use everyday — if we have had positive experiences dealing with another person or business in the past, we are likely to believe that future interactions will also be successful. Unfortunately, personal experience has obvious limitations: one's circle of acquaintances can only be so large, and personal experience helps very little when dealing with someone new.

When personal experience is insufficient, we can instead use a potential partner's *reputation*

[1] As suggested by (8), we currently do not try to assess motivation — poor performance caused by dishonesty or malice is indistinguishable from mere incompetence. Since the end result is the same, we do not think it is necessary to treat the sources of unacceptable performance separately. However, for convenience, we refer to behavior that contributes to the success of a transaction as "honest" and to behavior that does not as "dishonest."

[2] For simplicity, we assume only two individuals in each interaction. However, the trust problem and its solutions can be readily generalized to multi-party interactions.

to make the decision to interact or not. Simply put, a reputation is an evaluation of a individual's trustworthiness made by one's peers. In addition to other individuals that we know, these peers may also include trusted public entities and resources, such as published reviews and government ratings. Essentially, when we examine someone's reputation, we are using the opinions of those we know and trust to assess the risk of dealing with a stranger.

These two mechanism are generally sufficient for most real-world interactions. On-line, however, we are often confronted with the opportunity to interact with someone about whom no one we know has any experience or opinion. Because of the ease with which anonymous individuals can interact with people around the world, it is also unlikely there is any public information on these potential partners. We appear to be at a dead end: how can we can make a trust decision when there is no trusted information regarding these potential partners?

Of course, a safe strategy is to not trust strangers about whom we have no information. The obvious downside of this strategy is that we limit ourselves to a very small subset of the vast opportunities available on global networks. A second, somewhat riskier, approach is to use potentially unreliable sources to assist in making trust decisions. This approach is the essence of the reputation system: it uses the opinions of strangers to evaluate the trustworthiness of other strangers. Unlike using the opinions of peers, we have little reason to expect that strangers' evaluations of others are accurate, and have almost no recourse against such inaccuracies. Nevertheless, reputation systems have been successfully deployed in a wide range of on-line and distributed applications and are by and large successful.

Reputation system may seem to be creating trust out of thin air, but in reality they are simply exploiting the bold, but often true, assumption that the majority of users are mostly honest. In a properly designed reputation system, there should be no incentive for giving an otherwise unknown partner an inaccurate evaluation. In a reputation system with a suitably high participation rate, a user can expect to get a reasonably accurate estimate of a potential partner's trustworthiness, provided that partner has been active long enough to accumulate a reputation history. As is so often the case, the devil is in the details: what exactly does a "properly designed" reputation system entail? How difficult is it to build such a system? Under what circumstances will a user give inaccurate

evaluations and how can we detect them? We look at these questions in more detail below and in the remainder of this thesis.

1.1.1 The eBay Feedback Forum: An Example Reputation System

Perhaps the most widely used reputation system is the "Feedback Forum" of eBay (32), a major peer-to-peer market operator. While not without flaws — many of which we discuss in this thesis — the Feedback Forum proves largely effective at building trust among eBay users. eBay itself attributes at least some of the success of their market to the Feedback Forum. (36) The Feedback Forum is a widely studied and well understood example of an on-line reputation system that we use as a baseline when evaluating the novel systems we present in later chapters.

After each transaction, eBay encourages both the buyer and seller to leave feedback for each other. Feedback can be *positive*, *negative*, or *neutral*, although in practice neutral feedback is both very rare and generally considered a weak negative. (95) In addition to the quantitative feedback, users are asked to leave a short text description of the transaction. In May 2007, eBay introduced "Detailed Seller Ratings," a more comprehensive feedback system that permits buyers to leave more detailed quantitative feedback about sellers. (102; 31) At the time of this writing, there is too little information about this new feedback system to offer any conclusions regarding its effectiveness.

Users are limited to giving only one feedback per transaction, although engaging in repeated interactions allows one to leave multiple feedbacks for the same partner. However, the maximum amount of feedback a single user can contribute to a recipient's aggregated reputation statistics is limited. The feedback from each user is first summed, then limited to the range $[-1, 1]$. For example, if user A has left five feedbacks for user B, 3 positive, a neutral, and 1 negative, A's contribution to B's aggregated reputation scores will be treated as a single positive. This mechanism prevents trivial collusion and sybil attacks, where one user leaves many feedbacks for the same person in order to manipulate the target's reputation. While the contribution of feedback to the aggregated reputation statistics is limited, all feedback is recorded and available for inspection.

eBay provides several trust signals to assist members in making the decision to interact or not. Auction pages prominently display two aggregated summary statistics: the "Feedback Score," the

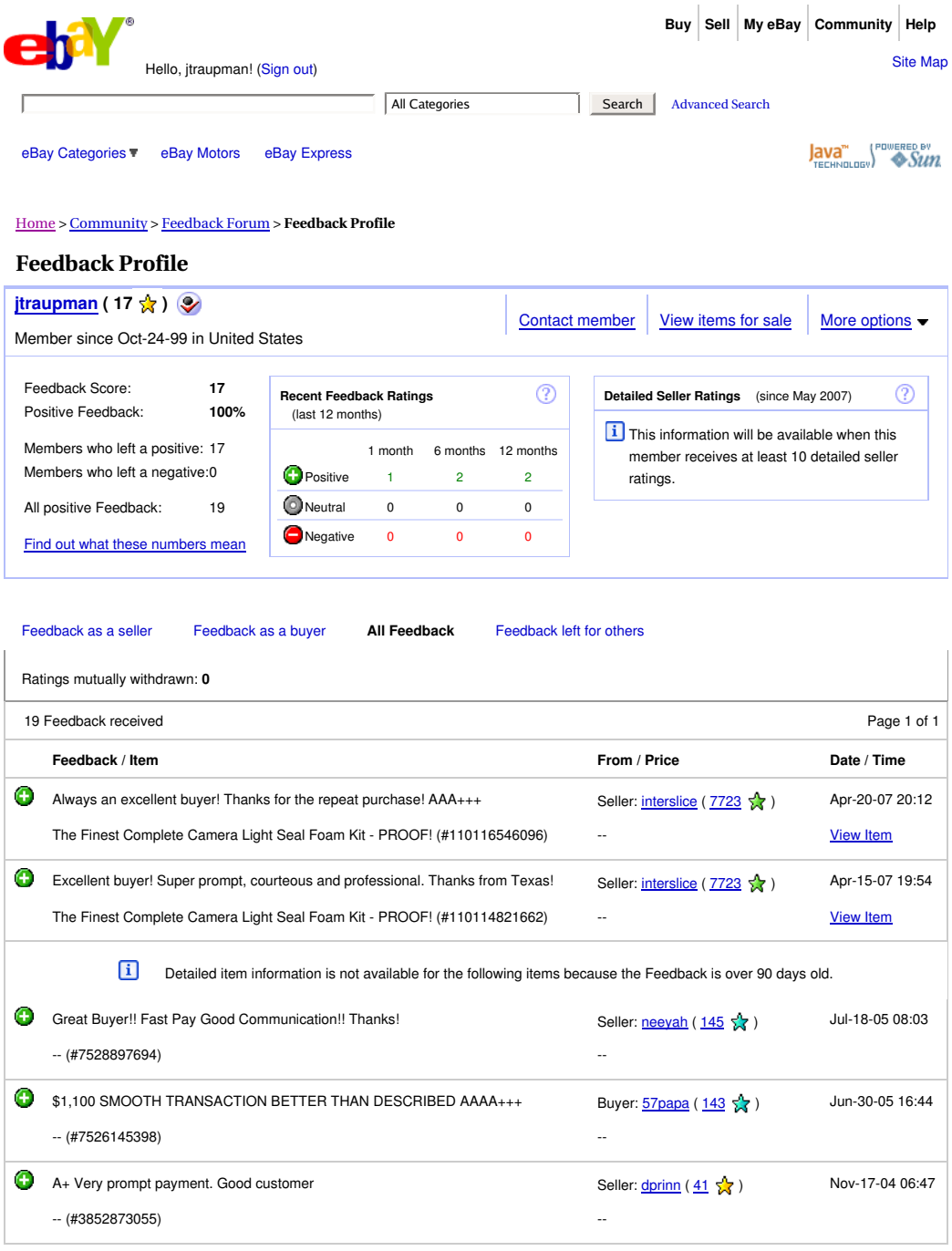


Figure 1.1: Snapshot of the author's eBay feedback summary page.

number of positives minus the number of negatives received, and the "Percent Positive Feedback," the percentage of total feedback that is positive. In addition, there are several icons that may be displayed next to a user's name. Colored stars provide a graphical representation of the feedback score, while other icons indicate that the member is a new user, has recently changed identities, has been ID verified, and/or is a high volume seller.

A member's feedback profile, accessed by clicking the feedback score, provides more detailed information. Figure 1.1 shows the author's feedback profile as an example. On the feedback profile page, a potential trading partner can find the same aggregated information as on the auction page as well as a breakdown of feedback received over time and all of a user's individual feedbacks with comments. Two notable problems with the feedback summary are that it is impossible to limit the listing to display only positive or only negative feedback and that detailed item information is only available for feedback left in the past 90 days. The former makes it very difficult to find only problematic feedback for high-volume sellers, despite the fact that the occasional complaint may reveal more trust information than a mountain of praise, if only because it less common. The lack of item details after 90 days is one of the features that makes sybil attacks (see Chapter 5) possible.

We examine specific features of the Feedback Forum in more detail in later chapters as we analyze its vulnerabilities to various attacks.

1.1.2 Reputation System Characteristics

Because we must make trust decisions in a wide range of endeavors, the domain of applications for reputation systems is naturally large. However, all reputation systems share a few common features and most domain-specific applications simply refine these traits to handle specific requirements. Reputation systems perform three broad tasks:

1. **Collecting feedback.** Individual users submit their opinions on others in the system.

2. **Computing reputations.** The collected feedback is aggregated to form reputation scores for each user.

3. **Assisting with trust decisions.** The reputations are presented to users in a way that assists them with making trust decisions.

The way in which these three operations are implemented varies across system.

Collecting Feedback

During feedback collection, the reputation system solicits opinions about user trustworthiness. In some systems, giving feedback is strictly optional and at the discretion of the individual users. Other systems require that feedback be left. Still others do not require feedback, but provide rewards to users that leave feedback. Requiring or rewarding feedback can increase the participation in the system, but it must be balanced against the increased risk of collecting dishonest or inaccurate feedback submitted by users whose only desire is to collect an incentive. (43; 73)

Another major design decision during the collection phase is the definition of the set of users permitted to give and receive feedback. In the most open systems, any user can evaluate any other user. This approach was used in the original version of eBay's Feedback Forum, but was discontinued after a few well publicized cases of abuse. (16; 33; 83) Other systems allow ratings of only a certain subset of users, such as retailer review sites that permit customers to give feedback about on-line stores registered with the site. (5; 35) Finally, some systems (like the current version of the eBay Feedback Forum) permit feedback to be given only in the context a transaction. Of these systems, some are *bidirectional* and allow both parties to leave feedback for each other. Others are *unidirectional* and solicit feedback from only one side, such as only recording feedback by buyers about sellers. Even in unidirectional systems, it may be possible for two users to both leave feedback about each other if they engage in two different transaction with their roles reversed in the second one.

The type of information collected also tends to differ across reputation systems in different domains. The simplest systems collect only a single bit of information: e.g. good/bad or successful/no information. Other systems allow both positive and negative feedback or positive/neutral/negative feedback in addition to a "no answer" category. Most of the systems we study use this approach. Finally, it is also possible to have finer gradations (e.g. 1-10 or even real values). In addition to nu-

meric ratings, many systems also ask for textual reviews or descriptions of user experiences. While such information can be very valuable to end users, it is much more difficult to aggregate and is typically just presented to queriers unprocessed.

In some cases, the application determines the type of feedback collected — for example, in peer-to-peer file sharing systems (systems used for trading files among individuals, often illegally), a user requesting a service can only reward a peer for fulfilling the request because it is impossible to tell whether peers that do not offer the requested file are deliberately withholding it or simply do not have it available. In other applications, the choice of scale is left up to the designer. Finer granularity allows the system to discriminate more finely, but may confuse users or make aggregation more difficult.

Related to the rating scale is choice of the number of dimensions of feedback to collect. Some reputation systems collect only one dimensional feedback (presumably on a trustworthy/untrustworthy continuum). Others break down different aspect of the user experience (such as item quality, customer service, packing and shipping for an on-line seller) and request separate ratings on each. Once again, collecting several orthogonal ratings may lead to more specific reputations, but runs the risk of confusing or annoying users, reducing the participation in the system. Often, systems will collect both an overall rating as well as optional, more specific evaluations of individual quality dimensions. (102)

While all reputation systems require some sort of feedback collection process, implementors have considerable freedom to choose different approaches based on the requirements of their specific domains. These decisions influence the type and quality of information available to the next phase of the reputation system, feedback aggregation. Individual designers need to recognize the trade-offs implicit in the choice of how to collect feedback and choose the approach most appropriate to their application.

Aggregating Feedback

Once enough feedback about a particular user has been collected, we can process this raw data to compute the user's reputation. Much of the research conducting in the reputation systems field concerns itself with developing aggregation strategies.

The simplest approach is to do nothing: collect the individual feedback ratings, along with optional textual reviews, and present this data unprocessed to the user. Such a system is trivially easy to create, but places all the burden of evaluating user trustworthiness on the querier. In large networks, the number of feedbacks per user can grow very large, making a manual assessment almost impossible. Furthermore, this approach does not eliminate the need for aggregation, it merely pushes to the edge of the network. Individual users still require some sort of algorithm (however informal) for distilling raw feedback into a metric that assists in making a trust decision.

Only slightly more complicated than doing nothing is to present the user with some simple statistics calculated from the raw feedback. Nearly all deployed systems calculate the mean feedback score or a similar metric, like the Percent Positive Feedback score used by eBay. Systems may also provide secondary information, like the standard deviation or the count of raw feedback, that can assist users in making estimates of confidence in the reputation. Systems that collect several dimensions of feedback data may calculate statistics both for the individual feedback dimensions as well as an overall aggregate.

Despite the simplicity of many deployed reputation systems' aggregation stage, we believe there are abundant opportunities to improve reputation systems by using more sophisticated techniques. In addition to the raw feedback data, other information, such as prior knowledge, side data (e.g. item price in a market, or file type in a file sharing network), interaction graph structure, changes in user interaction or feedback behavior over time, and many other sources can be used to improve the usefulness of the reputation system. In this thesis we are primarily concerned with developing aggregation techniques that make the reputation system more robust to missing data, inaccurate evaluations, and deliberate misuse.

Assisting the Trust Decision

Once the reputations have been completed, they can be used to help make trust decisions. In almost all current systems, this process is pushed onto the user. The reputation system returns aggregated reputations (and/or raw feedback data) in response to a user query and it is up to the querier to use this information to help make trust decisions. However, one could imagine that the system could do much more: by incorporating information about the particular trust decision being made (i.e. the specific risks and rewards involved) it could potentially return a tailored recommendation rather than a generic reputation.

In general, the line between algorithms for aggregating feedback to form reputations and ones for using reputations to make trust recommendations is somewhat blurry. For example, the perfect reputation system of Section 1.1 could potentially just return a single bit indicating whether or not to trust a specific partner. Is that single bit the reputation, or is the reputation some hidden value used to compute the recommendation signal?

To draw a clear distinction between these phases, we consider aggregation to be the process of creating reputations that are, in some sense, generic. They will, of course, be specific to the individual users and possibly also to individual queriers (as in the asymmetric reputations of (21)). However, the reputations created by aggregation contain no knowledge of the type of decision where they will be applied. We consider trust decision support to be the process of further refining reputations in the context of a particular potential interaction to yield an accept/reject signal for that transaction.

In the reputation systems we study, we assume only minimal decision support: they provide users with aggregated reputations, which they can use with their own individual decision logic. In the simulations we conduct, we look at only very simple decision procedures, such as soft and hard thresholds. Nevertheless, there appears to be many opportunities for improving both the accuracy and usability of reputation systems by making improvements to this phase of the process, but we must leave such investigations for future study.

1.1.3 Reputation System Applications

While the majority of reputation systems implement these three phases of operation, there are considerable differences between them depending on the particular domain of application. Because of the prevalence of trust decisions in interactions between people or organizations, reputation systems have found use in a variety of different areas.

Peer-to-peer Markets

Perhaps the most widely known reputation system application is the peer-to-peer markets. In a peer-to-peer market, the buyers and sellers are simply individual users of the marketplace site, in contrast to more familiar models of commerce where sellers tend to be large, persistent, independent entities. Perhaps the most familiar peer-to-peer market is the auction site eBay (32), where 1.9 billion items changed hands in 2005. (36)

As in most commercial transactions, exchanges in peer-to-peer markets involve risks for both the buyers and sellers. These risks are compounded by the global and more or less anonymous nature of interactions compared to more traditional retail models. The risks to the buyer are that the seller will accept payment without sending the purchased goods or that the item will be different than described on the sale page[3]. Sellers have the opposite problem: they risk sending a valuable (possibly irreplaceable) item to an unknown seller who may have sent a fraudulent payment.

Reputation systems for peer-to-peer markets must try to mitigate these risks. Because both sides assume risk, most systems use bidirectional feedback. Exchanges in peer-to-peer markets involve real monetary value, so a good reputation may also have value. (95; 7; 12; 27) Therefore, reputation systems for this application need to be particularly resistant to fraud committed against the reputation system itself in order remain effective.

Because reputation systems for peer-to-peer markets is the specific focus of this thesis, we provide more details of this application domain in Section 1.2.

[3]Many of the items for sale in peer-to-peer markets are used so inflated descriptions of item condition are fairly common

File Sharing Networks

Peer-to-peer file sharing systems provide another application for reputation systems that shares some similarities with marketplaces. Like peer-to-peer markets, file sharing involves an exchange between two individuals. As the name suggests, the items exchanged are files, which are typically large.

However, there are some key differences between file sharing and marketplaces. In most file sharing networks, the exchange is unidirectional: the "client" user requests a file that is sent by the "server" user, but provides nothing in return. The social contract in place in most file sharing systems is that users spend time acting as both clients and servers, so that the bandwidth and storage costs are distributed fairly among all users.

Anonymity is even more prevalent than in peer-to-peer markets, since the "killer app" of most file sharing networks is the illegal distribution of copyrighted media. Persistent pseudonyms are possible, but since physical addresses, credit cards, and other identifying information are not required to participate in these systems, it is very difficult to connect a file sharing user to a real world person.

There are two main risks in file sharing networks. The first is that users will "freeload" or act as clients but never as servers. (1; 61) Because sharing a file with a freeloading client requires the same bandwidth and storage cost as a fully participating client does, if too many freeloaders join the network the total burden of serving files increases without any benefit to the users willing to share. Some systems, notably BitTorrent (24), rely on a tit-for-tat scheme that forces users to both upload and download. However, since this approach is implemented in the network client software, it is possible for sophisticated freeloaders to engineer ways around it. (67)

Other systems either rely on reputation systems to rate users based on their willingness to share. (68; 41) Users can then refuse download requests from others who do not share frequently enough. In this sort of reputation system, feedback is generally unidirectional — downloaders can give a positive feedback to a user that has served a file, but it does not make much sense for servers to give their clients any feedback, since the clients provide nothing to the servers. Similarly, feedback is often limited to just a binary positive/no information indicator. Since it is generally impossible

to assess motivation for *not* sharing a file — a user may be refusing to share or may simply not have the file available — there is little reason to leave negative feedback for users who do not offer a specific file for download. In such a system users will only share files with others who have received a sufficiently high number of positive feedbacks, which indicate that they already shared some files.

Somewhat orthogonal to the problem of freeloading is the rise of content spam in file sharing networks. Since much of the content available on these systems is illegally copied, some rights holders employ the tactic of joining the network and offering what appears to be highly demanded content. However, upon retrieval, the downloader discovers that the files contain only a stern anti-piracy message. The spammers' goal with this tactic is to flood the network with such messages so that users give up and leave the network due to the time and bandwidth wasted downloading the spam files. The Credence system (111) offers a novel solution to this problem; in essence, using a collaborative filtering engine as a reputation system. Credence is specifically targeted at the content spam problem and does nothing to address freeloading.

One final concern with reputation systems for file sharing networks is that they tend to be more fully distributed than peer-to-peer markets. As a consequence of the type of files being shared, these systems must be architected to avoid any centralized coordination that may be an easy target of legal action. Reputation systems for file sharing must therefore also be fully distributed, leading to further complications with collecting, aggregating, and evaluating feedback. Peer-to-peer markets, as legal businesses, often rely on centralized search and communication servers both for efficiency and to lock customers in to a particular provider. Therefore, the logical place for a reputation system is on the market's central servers, greatly simplifying its implementation and eliminating certain problems specific to fully distributed implementations.

While reputation systems for file sharing networks are a fascinating subject in their own right, we do not examine them in detail in this thesis. However, we believe that some of the techniques we develop are general enough to apply to both markets and file sharing systems equally effectively.

On-line Social Networks

In social networks, the object is to make connections with other users, not necessarily a commercial or data exchange. Many sites, such as Friendster (46), MySpace (90), and Facebook (119), serve a purely social purpose. Others, like LinkedIn.com (79), exist to create and maintain professional connections. Beyond the on-line world, there are a myriad of social networks formed by our day-to-day personal and professional interactions, many of which can be be used when making trust decisions.

In all cases, the sites operate by encouraging users to create profiles describing themselves and their interests and then link to the profiles of friends in the network. Implicit in the concept of linking to friends is a trust relationship — users only link to others with whom they share interests. The goal of these systems is to create a concrete realization of a "small world" network, a graph with sparse connections that nonetheless has a small mean connection distance, the average number of links one must follow to connect any two nodes. (114) Small world networks have been observed within a wide range of natural and social processes, from graphs of academic collaborators (48; 89; 94; 51) to chirping crickets. (114)

Unlike markets and file sharing networks, the reputation system is not implemented on top of a social network. Instead, the social network application itself is the reputation system that users can query to find other trustworthy people for social interaction.

As a consequence of their design, most social networks use small-world approaches to making trust decisions. While difference schemes for mining trust information from social networks have been proposed (52; 58; 113) most operate by calculating some metric representing the connectivity between a user and a potential new social contact. Many also offer the ability to browse or search the social graph for others with similar interests.

Like file-sharing systems, social networks typically allow only binary positive/none feedbacks to be given in the form of friend links to other users. While there is no reason a "dislike" link or a more expressive rating scale would be incompatible with this application, we do not know of any social networks that use these reputation system features. The largest risk to social networks is from colluding users polluting the graph with phony links in a manner analogous to content spam in file

sharing systems. It seems likely that systems like Credence could be adapted to work in similar fashion in this domain.

In addition to finding new social partners, these networks can also be used as a supplemental reputation system for other applications. For example (53) use a simple social network and small-world techniques for filtering spam from legitimate messages.

We do not specifically study social networks in depth in this thesis. The small-world methods that most employ can be effective for evaluating reputations of strangers that are close to one's peer group, but are less useful when dealing with more distant users. Furthermore, the large scale and global search facilities of most on-line peer-to-peer markets lead to interaction graphs that do not exhibit small-world structure. However, several of the graph-based techniques presented in Chapter 5 borrow some technology from the approaches used in this domain and may be applicable to it.

Ad Hoc Networks

Ad hoc networks consist of a collection of nodes each equipped with a communication interface, typically a low powered radio. They are designed to dynamically reconfigure their connections with other nodes to perform computations, send messages, or report sensor readings. In most applications, efficient power use and robustness against changing network conditions and attacks by adversaries are of paramount importance. (14) gives a good survey of the important results in this field.

Reputation systems can help ad hoc networks achieve their resilience goals by detecting and isolating problem nodes. Unlike the other applications discussed so far, the actors in an ad hoc network are typically machines, not humans. Of course, the behavior of the nodes may be controlled by a person; in particular, nodes may be compromised by an attacker in attempt to block or modify traffic on the network.

Reputation systems for ad hoc networks typically work by exploiting the network's graph structure through the use of small-world techniques like social networks. When deciding which of several neighbors to use to forward a packet, a node will use its own personal experience with the nodes

in question along with the opinions of its neighbors and more distant nodes in the network. Most of the differences between ad hoc network reputation systems revolve around the way in which this feedback is aggregated. The decision of whether to trust a node to forward a packet is a trade-off between the perceived reliability of the node, as given by the reputation system, against other factors such as the power required or the available bandwidth of the node.

(13; 85; 86; 57) all present novel and interesting proposals for reputation systems in ad hoc networks. Like the other application discussed here, we look to such systems for inspiration but do not concentrate our investigations on this domain.

Other Examples

The four applications just presented represent only a few example of the applications of reputation systems. Essentially any distributed application that involves interactions and trust among peers can benefit from reputations. For example, reputation systems have been applied to the problem of determining which developers to trust in open source software projects (4) and to the filtering of newsgroup-like forums. (100) In practical use, the boundary between reputation systems, recommender systems, and social networks often blur as all attempt to distill global evaluations of quality from local opinions. Table 1.1 presents a number of well known reputation systems and summarizes their features and implementation.

For the remainder of this thesis, we concentrate our investigation on reputation systems for peer-to-peer markets. We choose this application domain for several reasons:

- Peer-to-peer markets are widely used, so improved reputation systems will have a direct benefit.
- The measurable value of the items exchanged creates strong incentives for adversaries to attempt to subvert the reputation system, so techniques for improving resilience to these attacks are needed.
- Reputation systems for peer-to-peer markets have been widely studied, yet the field still has many open problems.

Name	Application	Feedback type	Feedback scale	Aggregation technique
eBay Feedback Forum	Peer-to-peer market	bidirectional	positive/none/negative	Percent positive feedback Positive feedback count
Amazon Marketplace	Peer-to-peer market	unidirectional	integers 0–5	mean score
EigenTrust	File sharing	unspecified	real between -1 and 1	variation of PageRank
Credence	File sharing	unidirectional	binary positive/negative	weighted mean
LinkedIn	Social Network	bidirectional	binary link/no link	browsing and search only
TrustMail	Email filtering	bidirectional	1–10	recursive weighted mean
Advogato	Collaborative development	bidirectional	binary trust/no trust	Trust flow
CONFIDANT	Ad hoc networks	bidirectional	binary, two dimensions	Bayesian update

Table 1.1: Several popular reputation systems and their characteristics.

- Peer-to-peer markets avoid the ethical and legal controversy surrounding file sharing, another widely studied domain.

Nevertheless, most of the problems we examine and the solutions we develop do not strongly depend on assumptions valid only in peer-to-peer markets. It is our hope that many of these techniques will find application in other reputation system domains with only minor modification.

1.2 On-line Peer-to-peer Markets

As briefly discussed above, peer-to-peer markets are marketplaces where buyers and sellers come together to trade. In contrast with traditional retail models, sellers tend to be individuals or small businesses and have a less permanent presence in the market than a large retailer might. They also operate within the confines of the market rather than through a separate, independent site. However, this boundary is rather fuzzy — traditional retail businesses are playing an increasing role as sellers in on-line peer-to-peer markets (116; 98; 2) and the markets themselves have reacted by offering services that allow retailers to distinguish their virtual store from ordinary sellers. (110)

Peer-to-peer markets are nothing new; in a certain sense they are the most ancient form of commerce. We are surrounded by examples of peer-to-peer markets in the off-line world: classified ads, flea markets, bazaars, and other forums where individual buyers and sellers meet to trade. However, moving these markets on-line causes some fundamental changes in their operation. In traditional, off-line peer-to-peer markets, the buyer and seller nearly always meet face-to-face when exchanging goods for money. While they may not know each other beforehand nor have any means of following up after the sale, this in-person trading process mitigates much of the risk. Buyers can examine the goods for condition, suitability, and authenticity before paying, while sellers can insist on easily verified forms of payment like cash or a certified check from a local bank. In on-line peer-to-peer markets there is almost no opportunity for such a face-to-face exchange since the buyer and seller may hail from different cities or even nations.

It is something of a testament to basic human honesty (or possibly high tolerance to risk) that on-line peer-to-peer markets can operate as successfully as they do. Simply stated, sending money to

a anonymous stranger for an item advertised on the Internet seems like a very bad idea. Reputation systems and other protections offered by the market owner and payment service help mitigate some of the risk, but to be successful, on-line peer-to-peer markets must assume that the majority of their users are legitimate.

1.2.1 Mechanisms of On-line Peer-to-peer markets

Like reputation systems, on-line peer-to-peer markets share certain fundamental similarities but have many differences in their implementations. The majority of markets follow the same main steps to complete a transaction:

1. Sellers post advertisements for the goods they are selling.[4]
2. Buyers find goods they want to buy.
3. Buyer and seller negotiate a price.
4. Buyer provides payment to the seller.
5. Seller furnishes the item to the buyer.

After the transaction is finished, markets that implement a reputation system solicit feedback from one or both parties.

The first step, sellers advertising their products, is mostly largely the same across all on-line peer-to-peer markets. After all, without some information about what goods are for sale, the remaining steps cannot occur. The main difference among markets is that in some markets sellers are able to set up a virtual store that groups all of their for sale items together and often permits the seller to customize the look and feel of the site. In other systems, sellers must post individual ads that are then grouped by category. For an off-line example of this distinction, consider newspaper advertisements for used cars. Large auto dealerships may purchase a single large ad describing many vehicles they have for for sale. Smaller lots and individuals simply purchase classified space

[4] In some markets, the process is reversed: buyers advertise what they want and sellers offer goods for sale. For examples, this is the scheme used by "want to buy" classified ads and by the failed reverse auction startup iWant.com. While somewhat popular in business-to-business procurement, this model is not prevalent in the business-to-consumer and consumer-to-consumer markets on which we focus.

for each vehicle in the automobile section. Markets also differ in the type of detail they permit in their advertisements. Some allow only short, text ads like a newspaper classified, while others permit elaborate layout with text, photos, and HTML content.

The second step also differs very little from site to site. Most allow users to browse ads by categories such as type of item, date the ad was posted, price, location of the seller, and so on. The flexibility of the on-line medium allows the user to choose the most important categorization, unlike a newspaper or flea market, which must be organized only once. Nearly all on-line markets also provide a search service for seeking out specific items and may allow filtering of the search by some or all of the browsable categories. Some sites even allow search of historical data, such as ads for items already sold, permitting the buyer to investigate pricing trends.

The resolution of pricing is where the most substantial differences between markets arise. The simplest pricing scheme is to have the seller set a fixed price, as happens in a traditional retail store. The buyer then decides whether or not to accept the requested price. Some markets permit bargaining along the lines of "best offer" pricing. The buyer make an offer to pay a certain amount for the item (typically less than the asking price) and the seller decides whether or not to accept the offer, perhaps after receiving bids from several buyers. The market may automate this process, or it may just permit the bargaining process to occur via some other communication channel.

Slightly more complicated is the on-line auction, as pioneered by eBay. In an on-line auction, sellers list items for a fixed period of time, during which time buyers submit bids. At the end of auction period, the highest bidder wins the item. Auctions are another ancient tool of commerce adapted to the new frontier of the Internet, and there is a wide literature surrounding them (see (87) and (108) for an overview). The particular style of auction implemented by eBay is that of an ascending price (or "English") auction with proxy bidding: bidders submit their maximum bid and the proxy automatically increases the bids until only the maximum bidder remains. Under a private value assumption, this type of auction is equivalent to a sealed bid, second price auction.

On-line auctions are a rich field of study in their own right, with many unanswered questions about buyer and seller strategies, the role of specific auction features, and generalizations such as multi-object auctions. A good summary of the state of the field can be found in (92).

Site name	Advertiser	Pricing scheme(s)	Involved in payment	Involved in delivery	Reputation system
eBay	Seller	Auction Fixed price	Optional	No	Yes
Amazon Marketplace	Seller	Fixed price	Yes	Yes	Yes
Craigslist.org	Mostly sellers	Fixed price Negotiable	No	No	No
Yahoo! Auctions	Seller	Auction Fixed price	No	No	Yes
Classifieds.com	Mostly sellers	Fixed price Negotiable	No	No	No
eWanted.com	Buyers	Fixed price Negotiable	No	No	Yes

Table 1.2: Characteristics of several popular on-line peer-to-peer marketplaces.

The final two steps are generally implemented in one of two ways: either the market intermediates the exchange of goods for payment or they do not. If the market participates in the exchange, then payment is typically through the host site's "shopping cart" interface and shipment is arranged by the market. If the market does not participate, then the buyer and seller must negotiate the timing and method of payment and shipment. Not surprisingly, there are markets that blur this distinction, such as eBay, which owns the PayPal.com payment processing service, and thus interjects itself during the payment phase if both buyer and seller agree to use the PayPal service.

Table 1.2 summarizes the characteristics of several on-line peer-to-peer markets. Some of these implementation choices can influence the type of reputation system used. For example, Amazon marketplace implements unidirectional feedback where buyers can rate sellers but not the other way around. Because the market intermediates the purchase process, there is little need for feedback about buyers. While all of the reputation systems we consider in this thesis assume bidirectional feedback, we do not make any further assumptions about the details of the market's implementation.

1.2.2 Trust in Peer-to-peer Markets

As discussed in Section 1.1, on-line peer-to-peer markets involve trust decision, much like all commercial transactions. The buyer must trust the seller to deliver an item of the type and condition

advertised, while the seller must trust the buyer to make a prompt and legitimate payment. However, there are several specific trust issues that confront on-line peer-to-peer markets. Some are inherent in the nature of the application while others arise out of common practices in these markets. Both need to be addressed by any system claiming to assist with making trust decisions in this domain.

Non-persistent Pseudonyms

The largest single challenge to trust in on-line markets is the essential anonymity of buyers and sellers. In most markets, the only information necessary to create a user account is a valid email address. Some also require a credit card, particularly for billing seller fees, but neither is hard to obtain nor are they reliable forms of identification. Some markets, notably Craigslist.org, provide services that deliberatively anonymize users' identities — you will almost never see the same anonymous user ID twice on Craigslist. In other markets, users are allowed to choose a pseudonymous user name that is semi-persistent: users are encouraged to maintain the same pseudonym, but there is little to prevent the acquisition of a new pseudonym or even multiple simultaneous identities.

Non-persistent pseudonyms have far reaching consequences for trust in peer-to-peer markets. Fundamentally, one can never tell whether or not a given user name is the same or different than one encountered earlier. In theory, this permits users to trade identities and thus reputation, but there is little evidence that these trades happen in practice. Far more common is the problem of hijacked accounts: a criminal obtains the user name and password for a legitimate account holder (often through "phishing" or other social hacking attack), then uses the account and its good reputation to defraud buyers. (77; 101)

The fact that it is easy to switch identities also renders the concept of a "bad" reputation essentially useless in on-line markets. Anyone whose reputation falls below that of a new user can easily discard the old identity and register a new account, thus shedding the poor reputation (45; 42). New users are thus required to pay a form of "initiation dues," engaging in small transactions or building a reputation as a buyer before trying to sell, to prove their good intentions prior to being trusted by other users. (95) However, even these initiation dues are not enough to deter a determined attacker,

who can easily sell a pile of low cost items to win a good reputation prior to trashing it through a single large fraud.

Just as anonymity permits users from identifying their trading partners, it also prevents them from verifying the validity of feedback in the reputation system. By hiding behind easily created pseudonyms, it is possible for an attacker to boost his own reputation through fake transactions with his other pseudonyms (the so-called sybil attack) or to join with other attackers to increase each others' reputation (a collusion attack).

Despite the many problems introduced by anonymity, it appears to be a fact that we must accept. There have been proposals for both identity verification systems and persistent pseudonyms (45), yet neither has achieved widespread adoption. One possible solution is to make creating accounts and conducting transactions more difficult or costly, but such a scheme must be balanced against the risk of driving away potential customers. (9) Given the essentially fluid nature of identity on-line, we do not envision that the anonymity problem can be "solved." Rather, we believe that successful reputations systems must regard non-persistent identities as a fixed assumption and do their best despite them.

Asymmetry of Risk

Buyers and sellers in peer-to-peer marketplaces do not equally assume the risks of conducting a transaction. Due to both the nature of the risks and marketplace conventions, the buyer typically is exposed to greater risk than the seller. While the nature of the risk is similar for both, namely that the other side does not complete the transaction, the seller can take some steps to mitigate the cost of failure that the buyer cannot. The primary risk to the seller is the buyer either not paying or paying with a fraudulent instrument. Conversely, the buyer risks sending payment for an unseen item that may either never arrive or is substantially different from its advertisement.

By convention, in most markets the seller does not ship an item until after receiving payment. There is no hard and fast regulation enforcing this rule and nothing to prevent a seller from offering credit to the buyer; however, it holds true in all the example markets we examined. Of course, this strategy allows the seller to mitigate a great deal of risk: if a buyer does not pay, the seller never

sends the item. Thus the costs incurred because of a non-paying buyer are minimal: listing fees to advertise the item again, the marginal cost of keeping the item in stock for a few days longer, and some of the seller's time trying to collect payment and then re-listing the item. Of course, sellers find non-paying buyers frustrating, but unless they are an endemic problem, the real costs are only slightly more than an annoyance.

Buyers on the other hand, are hurt by this convention, since they must send payment on faith that the item will arrive. In the conventional retail model, with stores that are mostly permanent fixtures of the community, a failure to deliver can be resolved through a variety of channels. However, in the on-line environment, the sellers must be presumed anonymous, so the buyer needs high confidence in the reputation system to take this type of risk.

Assessing the other side's behavior is also easier for sellers. There are only a handful of mechanisms through which a buyer can make payment. Some are more secure and easily verified (cash, bank checks, credit cards) than others (personal checks), and sellers are generally permitted to specify the terms of payment. Therefore, when a payment is received, the seller can quickly determine whether the payment is legitimate before sending the item.

The buyer has a harder task: the types of items for sale is effectively unlimited, and if selling used merchandise is permitted, the condition of the item may be difficult to determine from an on-line description and a handful of photos, not to mention deliberate overstatement of the condition by the seller[5]. So in addition to having to send payment in advance, the buyer has a more difficult task determining whether the item actually lives up to its description, assuming it arrives at all.

While one solution to this problem is a reliable reputation system, there are several other ways of reducing or equalizing the risks, which we do not explore in this thesis. Some auctions provide buyers with insurance against undelivered or misrepresented items — Amazon Marketplace protects buyers that pay through the Amazon.com checkout interface, eBay offers different levels of protection for items paid through their PayPal service.

Another solution is to keep goods, payment, or both in escrow until both buyer and seller are satisfied with the transaction. On-line escrow services, such as escrow.com (39), can reduce both

[5] One occasionally encounters descriptions along the lines of "mint+," i.e. better than new.

parties' risk by acting as a neutral third intermediary that verifies payment but does not release it to the seller until the buyer has had time to inspect the item and determine whether or not it is satisfactory. Because escrow adds additional time and fees to the transaction, it is used most often with larger denomination sales.

Primacy of Reputations

In most markets, the reputation system is the primary means of mitigating risk. Some markets do provide some form of insurance against failed transactions, but market owners have their own incentives to avoid risk — indemnifying their users will almost certainly increase the cost of operating the market, a cost the will eventually be passed to the users. Likewise, some markets have mechanisms in place to confirm identities or prevent arbitrary creation of pseudonyms, but a truly reliable identity guarantee has not been feasible to date.

Therefore, reputation systems play a key role in building trust in on-line peer-to-peer markets. A properly constructed reputation system assists both buyers and sellers in managing their risk through a number of different processes:

1. Accurate reputations assist in making trust decisions. Obviously, if a reputation reliably predicts a user's future behavior, others can use the reputation to gauge the risk of interaction.

2. Reputations can encourage persistent identities. Because a good reputation is valuable, a user who has spent the effort to build a good reputation has an incentive to keep the same identity in order to continue to benefit from the reputation.

3. Reputations can stimulate good behavior. If a good reputation is difficult to build up but easy to tear down — the so-called "initiation dues" and "stoning" of (95) — then users will need to engage in consistent honest behavior in order to remain active in the market.

Furthermore, a reputation system can be designed to counteract some of the risk asymmetries present in this application.

Because reputations are often the main (or only) source of trust information in on-line markets, it is important that reputation systems be designed well. Not only should the system achieve the

above goals, but it also must resist deliberate attempts at subversion. If reputations are valuable, as they must be to be effective, then someone will try to find an easier route to a good reputation.

In addition to direct attacks on the reputation system, nefarious users will also try to subvert other marketplace processes for personal gain. An effective reputation system should permit the marketplace user community to police itself and greatly reduce the cost of this type of bad behavior.

Even simple reputation systems achieve many of these goals, as is evident by the success of peer-to-peer markets despite their risks. It is our belief that reputation systems designed to address the specific issues facing peer-to-peer markets will serve both the marketplaces and their users by increasing the ease with which reliable trust decisions can be made.

1.2.3 Challenges for Peer-to-peer Market Reputation Systems

Attacks on the reputation system are simply a fact of life — we have little hope of eliminating them. Instead, the reputation system must be designed with robustness against attackers always in mind. While we cannot stop the attacks, we can limit their effectiveness and influence on the system's performance. In this section we outline some of the attacks that the reputation system should prevent or mitigate. Designing reputation systems to be robust against attacks is the key challenge we address in this thesis.

Retaliation

Ideally, users would always leave accurate feedback that reflects the behavior of their partners. However, inaccurate feedback is almost certain to arise. Often it is unintentional, particularly in systems that have a complex feedback scale. Other times, it is deliberately misleading. The reputation system must be aware that feedback data may contain errors or deliberate misrepresentation and adapt accordingly.

Perhaps the most common form of inaccurate feedback is to leave multiple positives for an accomplice or multiple negatives for a competitor, in an attempt to build up or tear down the target's reputation. These so-called "ballot stuffing" and "bad mouthing" attacks can be mitigated to some

degree by only counting one feedback (typically either the first or the last one left) from each unique user account. We address these concerns below in the sections on collusion and sybil attacks.

Assume, for now, that users are not colluding and that the reputation system is a simple bidirectional system where users can leave either positive or negative feedback (or opt not to leave feedback). Under these conditions, inaccurate feedback means either leaving a positive for someone who was dishonest, or leaving a negative for an honest partner. Leaving a positive despite dishonest behavior is not very plausible. The user receives no benefit for leaving feedback (again, assuming there is no collusion) so certainly has no incentive to reward bad behavior. While difficult to prove, we believe that a positive feedback is a strong indicator that the transaction was completed successfully.

The other case, given a negative to an honest partner, happens quite frequently in the form of *retaliatory negative feedback*. In the typical retaliation scenario, one side of transaction (user A) behaves honestly, while A's partner (user B) does not. Both to punish user B and to warn B's future partners, A leaves B a negative feedback. B then sees this negative feedback and leaves a negative for A in retaliation, even though A completely abided by his half of the contract.

With retaliatory negative feedback, it is difficult to ascertain which negative is accurate, and should thus count against the user's reputation, and which is merely retaliation and should be ignored or discounted. Furthermore, retaliatory negatives have a chilling effect on participation in the feedback process. A single negative feedback has a large effect on the reputation of a user who has participated in only a few transactions, but scarcely any effect on users with long histories. Experienced users exploit this asymmetry by rarely leaving feedback first and always retaliating for received negatives, even ones that are justified. Users aware of this tactic often will not leave negative feedback except in the case of outright fraud out of fear of damage to their reputation. Retaliation is thus used as a threat by experienced users in order to discourage a partner from reporting bad behavior.

This tactic is another instance of the asymmetry of risk in present-day peer-to-peer markets. Because sellers, on average, engage in transactions more frequently than buyers, they tend to have longer feedback histories and are more likely to use retaliation as a threat. Despite the fact that a

seller can evaluate a buyer's performance even before sending the item, buyers are twice as likely to leave the first feedback in transactions where both ultimately leave feedback. (95) The end result is that negative feedback is discouraged and underreported, which damages the ability of users to rely on reputations to estimate partners' reliability. Some researchers believe that the market owners (specifically eBay) ignore this problem because a marketplace with a higher rate of positive feedback appears safer and more inviting to new customers. (16)

One frequently proposed solution to this problem is to implement the feedback process so that it is blind: a user cannot view the feedback he received from a partner until after he leaves his feedback. While we believe that blind feedback would be a major improvement for a very small effort, we do not think that it would be sufficient to solve all retaliation problems. Because there is typically out-of-band communication between buyer and seller about which the reputation system is unaware, users can often guess what type of feedback will be received and respond accordingly, in a form of "pre-emptive" retaliation.

One proposed solution that we emphatically dislike is eBay's recently implemented scheme of "mutual withdrawal." (34; 115) Under this policy, if both users agree to withdraw feedback for a transaction, then both feedbacks are erased from the system and are labeled withdrawn. With retaliatory negative feedback, at least *one* of the negatives is accurate, and browsing the feedback comments can often help the user determine which one. With withdrawn feedback, all that is known is that there was some sort of dispute (and most likely at least one negative), but it is impossible to tell which party is at fault. More importantly, withdrawn feedback is not counted toward the key aggregated scores (the Percent Positive Feedback and the Feedback Score). Unlike these scores, the number of withdrawn feedbacks is not displayed prominently on the main listing pages.

In Chapter 3, we examine the problem of retaliation from a game theoretic point of view and show that it is more than a mere nuisance. In our experiments, reputation systems that permitted retaliation are unable to maintain stable cooperation in the marketplace. In Chapter 4 we present EM-Trust, a reputation system that uses latent variable models to predict the most likely user honesties even with high rates of retaliation.

Sybil Attacks and Collusion

The ease with which individuals can create user accounts in peer-to-peer on-line markets and the essentially untraceable nature of the identities underlying these accounts lead to another broad class of attacks where groups of user accounts work together to manipulate the reputation system. In a *collusion attack*, a group of separate individuals conspire to commit some sort of reputation fraud. In a *sybil attack*, a single attacker creates many user accounts and uses these fake accounts to manipulate the system. The primary difference between the two types of attacks is that in a collusion attack, the users involved are all normal participants in the market who engage in normal transactions with other users. Aside from possibly faked transactions with co-conspirators, these user accounts look perfectly ordinary. In a sybil attack, the sybils' only purpose is to alter reputations. The two attacks can be combined: for example, a group of independent co-conspirators may also use sybils to build up each others' reputations.

The type of reputation fraud committed in these attacks tends to fall into two categories: "ballot stuffing" and "bad mouthing." (95) In a ballot stuffing attack, the goal is to increase the reputations of a particular set of users. For example, a group of users may all agree to leave each other a positive feedback or an attacker may create a collection of sybil accounts to leave one positive feedback each for his main identity. Under a simple aggregation scheme like the one used on eBay, these attacks allow arbitrary growth in reputation for a minimal investment. In a bad mouthing attack, the goal is reversed: the conspirators or sybils leave negative feedback in an attempt to lower another user's reputation. Typically, the aim of this type of attack is a to target a competitor or exact revenge against another user.

All forms of these attacks cut right to the heart of the reputation system by diminishing the value of a good reputation. Ordinarily, reputations require effort to build up and are easy to tear down if not maintained. This continual investment (by engaging in honest trading) gives reputations their value. However, an effective ballot stuffing attack short circuits the "initiation dues" process and allows a new user to quickly and easily build up a good reputation. Bad mouthing attacks provide a way to destroy the reputations of other users despite their consistent honesty.

A commonly implemented defense against these attacks is to limit feedback to only one per

user pair (i.e. only one feedback left by user A for user B is counted, usually the first or the last one). While this strategy can reduce the impact of ballot stuffing or bad mouthing by a single user, it does little to defend against groups of users or sybils. Imposing transaction and user creation fees (9) may allow markets to render these attacks unprofitable by making the cost of a marginal gain in reputation higher than its benefit. However, it is far from clear whether these fees can be set high enough to be effective while not measurably impacting legitimate transactions.

Collusion attacks of both types are, in general, very hard to prevent. Certain obvious configurations of conspirators, such as feedback cliques or a single group that conducts repeated bad mouthing attacks, are fairly visible. However, if a group of conspirators is large, has relatively sparse interactions within the group, and has members with a sufficiently high rate of interaction with out-group users, then detection will be very difficult. The essential problem of collusion detection is verifying the legitimacy of individual transactions. Since most peer-to-peer markets are not directly involved in the payment or shipping, it is almost impossible to distinguish between a real transaction and a cleverly obfuscated fake. In our opinion, collusion detection will require the intervention of trained investigators in addition to automated tools — reputation systems alone cannot be expected to detect illegitimate transactions using feedback alone.

We are more optimistic about detecting sybil attacks. While determining whether a user has multiple accounts is also very difficult, we can recognize sybil accounts that are used only to conduct reputation attacks (instead of, say, having a separate buying and selling account). Chapter 5 presents Relative Rank and RAW, two related systems for defending against sybil ballot stuffing attacks. Both systems use the graph structure formed by user feedback to down-weight feedback from nodes that appear to be sybils.

User Apathy

While not specifically an attack, low user participation can also have detrimental effects on reputation system performance. If too few users leave feedback, then the aggregated reputations will suffer due to lack of data. Furthermore, participation must be honest — leaving deliberately deceptive or random feedback is worse than not leaving feedback.

As mentioned above, retaliation deters users from participating by honest leaving negative feedback, leading to an overall positive bias in many reputation systems. Our EM-Trust system specifically addresses this problem by reducing the reputation penalty of retaliatory negative feedback. In our game theoretic studies of reputation systems (Chapter 3), we look in some detail at the problem of user apathy and how it affects market performance.

1.2.4 Non-reputation Attacks

In addition to attacks on the reputation system, markets must contend with other ways in which dishonest users may try to abuse market processes for their own ends. While certainly an interesting and relevant topic, these attacks are beyond the scope of this work. We encourage further exploitation of methods that can mitigate these problems, and believe that reputation systems will likely play a large role in defending against these attacks.

Shill Bidding

The practice of shill bidding is somewhat related to sybil attacks on reputation systems. In an auction-based market, if a seller is dissatisfied with the prices legitimate users are bidding, he may employ a shill, either another conspiring user or a sybil account created specifically for this purpose, to place bids on the item in an attempt to raise its selling price. In the commonly used second price auctions, shill bidders can increase the amount the winner must pay to nearly their maximum bid amount (in essence turning a second-price auction into a first-price one). (87, p. 17) Furthermore, because bidders in an English auction may readjust their valuations based on information learned through observing others' bids, shill bidding can server as a mechanism by which the seller can increase bidders valuations. (112) Counter to this intuition, there are also analyses that suggest that shill bidding is counterproductive and actually hurts the seller. (18; 72) Regardless of its benefits or lack thereof, shill bidding distorts the auction process and should be discouraged.

One challenge in detecting shill bidding through a reputation system is that, if successful, the shill bidder does not win the auction, so no feedback concerning the shill account is ever entered

into the reputation system. However, markets can easily retain bid information (even if it is not made public) and mine it to detect shills.

Bait and Switch

Another phenomenon that has made a successful transition from the off-line to the on-line world is the age old tactic of "bait and switch." An unscrupulous seller advertises one item, usually a desirable item for a very low price, then actually delivers something different. Often, the difference will be described, but only in very small type or in a hidden part of the advertisement.

We do not feel it is worthwhile to make a large distinction between this attack and the case of a seller who inflates descriptions of an item's condition. There may be a motivational difference (e.g. outright fraud versus incompetence or overzealousness), but the buyer suffers regardless. We believe that the same mechanisms for protected against ordinary transaction failure, such as well-designed reputation system, insurance, escrow, etc., will be equally effective against bait and switch.

There is also a fine line between illegal bait and switch and simply savvy marketing. In (38), Ellison and Ellison study the tactics of small, low margin retailers competing in the commodity PC parts business. They find that successful retailers attract customers to their sites by advertising very low prices on shopping search engines, then encouraging the purchase of more profitable items. For example, a retailer may offer a below cost price on a generic memory module without a warranty, but on the same page also link to a name brand, warrantied module that is a more profitable sale. Unlike illegal bait and switch, where the customer receives a product different than what was ordered, this obfuscation tactic is legitimate and not substantially different than "loss leaders" and similar sales approaches used by traditional retailers for years.

Hidden Costs

Another common tactic is for sellers to offer a great price on a desirable item, but with large hidden costs such as high handling charges or separate charges for normally bundled accessories. A widely publicized study (59) found that many on-line buyers do not account for shipping costs when comparing item prices from competing sellers.

When taken to an extreme, such practices are an attempt at fraud and should be handled by the reputation system or other defense mechanisms. In milder cases, though, one may argue that these hidden fees are merely good salesmanship and that *caveat emptor* holds as true on-line as it does in the real world.

1.3 Thesis Outline

Reputations systems for peer-to-peer on-line markets is a large field, and we have broadly examined many of its aspects in this introduction. Due to time and resource limitations, we cannot give equal attention to all problems in this domain. The remainder of the this thesis concentrates on one specific subtopic: designing reputation system techniques that are robust against attacks. In particular, we examine the effects of retaliation and user apathy on reputation system performance, propose strategies for limiting the detrimental effects of retaliation, and investigate techniques for defending against sybil attacks. We look at several different reputation systems, both existing and novel, a brief description of which can be found in Table 1.3.

The following chapter provides some background on this topic and summarizes the related work in this field. We also briefly discuss some existing proposals for reputation systems that time does not permit us to investigate directly.

Chapter 3 presents the results of a series of experiments exploring the effects of retaliation and user participation on reputation system performance. Because reputation systems are typically only one part of a defense in depth strategy, it is quite difficult to separate out just what the contribution the reputation system is making toward fostering cooperation in the market. By removing all external forces, we are able to study reputation systems' abilities to maintain cooperation in a market where agents evolve to maximize their performance under the conditions imposed by the system under test.

Chapter 4 presents the EM-Trust system, a reputation system designed to mitigate the influence of retaliatory negative feedback. By using a latent variable statistical model, EM-Trust is able to estimate the most likely distribution of blame in transactions where both sides leave a negative

	Name	Discussed in	Problems or Contributions
Existing Systems	Percent Positive Feedback	Chapters 3–5	Virtually no defenses against reputation fraud
	EigenTrust	Chapter 5	Claims of collusion resistance not true Not suitable for use in peer-to-peer markets
Novel Systems	Blind PPF	Chapter 3	Somewhat robust against retaliatory negative feedback
	EM-Trust	Chapter 4	Robust against retaliatory negative feedback
	Bayesian PPF	Chapter 4	Improves PPF's performance with sparse data
	Bayesian EM-Trust	Chapter 4	Robust against retaliation and sparse data
	Relative Rank	Chapter 5	Makes EigenTrust useable in peer-to-peer markets
	RAW	Chapter 5	Highly sybil resistant, personalizeable reputation system

Table 1.3: Reputation systems discussed in this thesis.

feedback. We also look at Bayesian techniques for minimizing the negative effects of the very sparse data that reputation systems must often use.

In Chapter 5, we discuss techniques for defending against sybil attacks. In particular, we present two algorithms, Relative Rank and RAW, that minimize the amount of reputation gain an attacker can realize using sybils. Both algorithms are extensions of the popular PageRank (93; 11) algorithm used in web search applications. Used together, RAW and Relative Ranks meet an important necessary condition for being sybilproof: they implement an asymmetric reputation system, where one's reputation depends on who is querying it. (21)

The final chapter contains some concluding remarks and suggestions of further work in this field. We also include an appendix describing the various versions of the marketplace simulator we use to evaluate these systems. Because real marketplaces are typically closed, proprietary systems, it proved impossible to test these approaches with real users. Instead, we developed a simulation package and calibrated its behavior with data from (95), an empirical study of user behavior at eBay.

Chapter 2

Background and Related Work

In this chapter, we present a summary of the work that serves as a background to this thesis. As discussed in Chapter 1, the problem of trust and reputation is ancient and there is undoubtedly much more to the various facets of this topic than what is discussed here. We begin by looking at non-computational studies of trust and reputation, primarily from the economics literature. We then look at early proposals for on-line marketplaces and reputations. Finally, we examine and critique current proposals for reputation systems.

2.1 Trust, Reputation, and Markets in Economics

Not surprisingly, the economics community has long held an interest in the notions of trust and reputation. Since most economic transactions involve both a need to cooperate and a risk to the participants, it is only natural that trust is a recurring theme in many branches of this subject.

2.1.1 Theoretical Studies of Reputation

An early exploration of the role of reputation is used to explain Selton's "Chain Store Paradox." (99) This game-theoretic problem pits a single monopolist against a group of independent, smaller competitors. In the classical formulation, the monopolist represents a large chain store that competes in multiple separate markets while the competitors are smaller companies that only operate in one

		Competitor	
		Enter	Decline
Monopolist	Fight	$b-1, -1$	$0, a$
	Acquiesce	$b, 0$	$0, a$

Table 2.1: Payoffs for (competitor, monopolist) in a single stage of the Chain Store game. In the formulation of (75) it is necessary that $a > 1$ and $0 < b < 1$, which is slightly different than the original payoffs in (99).

of the markets. The game proceeds in series of stages, in each of which a competitor must decide whether or not to enter the market. The monopolist must then decide whether to fight the competitor or acquiesce. The stage game payoffs are given in Table 2.1.

The paradox arises because the game has a subgame perfect equilibrium where the competitors always enter and the monopolist always acquiesces. Yet, this equilibrium does not explain the more plausible behavior where the monopolist fights in order to acquire a reputation as a tough competitor and competitors are scared away from entering the market. Stelten (99) proposes a non-game-theoretic decision theory to explain this behavior as a deviation from pure rationality. Both Milgrom and Roberts (88) and Kreps and Wilson (75) propose similar, game-theoretic explanations for the paradox in terms of imperfect information. If the monopolist's payoffs are not common knowledge, then it may cultivate a reputation for fighting and scare away the competitors. In the simplest formulation, the monopolist's payoffs may either be described as above or as an alternate formulation where it receives some positive value for fighting. Under this uncertainty, equilibria arise where the competitors decline to enter the market and the monopolist fights, at least for the majority of stages. Kreps and Wilson introduce the concept of *sequential equilibrium* (76), an equilibrium slightly less stringent than the well-known subgame-perfect equilibrium, to reason about strategies under these conditions.

The Chain Store game is very similar to another widely studied game, the Iterated Prisoners Dilemma (IPD), which also possesses paradoxical features. In this game, two players repeatedly play the Prisoners' Dilemma game, shown in Table 2.2, with each player receiving the sum of the stage payoffs as their final value. The finitely repeated Prisoners' Dilemma has a subgame-perfect equilibrium where both players defect at each stage, which seems to defy intuition about the best

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	r, r	s, t
	Defect	t, s	p, p

Table 2.2: The Prisoners' Dilemma in normal form. Actual payoff values must be such that $s < p < r < t$.

way to play the game. In an empirical study of the IPD, Axelrod (6) found that agents playing strategies like "tit-for-tat" (choose the same strategy as the opposing player chose in the previous stage) do better than ones playing the always-defect strategy in a tournament with agents playing a variety of strategies. Just like in the Chain Store game, the IPD appears to reward deviation from pure rationality. Kreps et al. (74) cast the IPD in terms of imperfect information and sequential equilibrium, using a similar approach to their work on the Chain Store paradox, and show that cooperation in this framework is a perfectly rational reaction to the other player's reputation.

John and Nachman (63) and Diamond (29) both study the role of reputation in debt markets. Using similar techniques to the above work on the Chain Store game and IPD, they formulate games that model the actions of firms seeking debt financing. They find that reputation effects help to eliminate the moral hazard problem of firms underinvestment after receiving financing. Firms that cultivate a reputation for sound investment are able to secure lower interest rates and thus higher profits. The two author's models are different, but both reach similar conclusions regarding the benefits of reputation for debt financing.

Chemmanur and Fulghieri (20) also explore the role of reputation in debt financing, but they concentrate on the reputation of the lender, not the borrower. In their model, banks can build a reputation for renegotiating the loans of borrowers in financial distress while the holders of publicly traded debt cannot. Such a reputation allows banks to charge a higher interest rate than bond issuers, but also causes high risk borrowers to prefer bank financing. Low risk firms prefer publicly traded debt because the lower interest rate allows greater profitability despite the risk of liquidation in the unlikely event of financial distress.

Harrington (65) uses a repeated two period game with overlapping generations to examine rep-

utation in a political context. Politicians in his model receive benefit from both implementing their party's ideology and from holding office, which often entails moderating the policy in order to attract more voters. In equilibrium, Harrington finds that politicians from both parties will adopt policies more moderate than their true ideologies, yet still separate from each other.

Mailath and Samuelson (81) study the effect of reputations, linked to brands, that can be traded. In their model, companies can build reputations based on whether they provide "competent" or "inept" service to consumers. Occasionally, a firm exits the market and sells its reputation to a new owner. Consumers are not able to observe the transfer of reputation and can only slowly update their assessment by observing firms' behavior over time. Inept firms prefer to purchase either very good reputations, which they can then exploit by trading on the good name of the previous owner for a few rounds, or poor reputations, which are cheap. Competent firms purchase average reputations; they can easily build the reputation through good service, so purchasing a good reputation initially is not worth the cost while poor reputations require too much time to polish. The authors also briefly examine the effects of various signaling mechanisms that allow firms to telegraph to consumers that a change in management has occurred.

Tadelis (103; 104) also studies reputation as a tradeable asset using a finite period game. He observes two conflicting effects that determine who buys good reputations. The "Reputation Maintenance Effect" arises because competent firms can more easily maintain a good reputation, so over a long enough period, they benefit more from initially buying and then maintaining a reputable name. The "Reputation Start-up Effect" is a consequence of the fact that inept agents cannot build up their own reputation, so if poor behavior does not tarnish a name quickly enough, it is profitable for an inept firm to purchase a good name and deplete it. These results are broadly aligned with those found in (81), namely that high reputations tend to be purchased by inept firms (the start-up effect outweighs the maintenance effect), while competent firms will purchase average reputations. One notable feature of Tadelis's model is that he also considers the possibility that firms may simply discard a bad reputation and start over with a new, previously unseen name, a behavior widely observed in on-line markets, unlike name trading.

Kennes and Schiff (69) look at reputation specifically within the context of a reputation system. In their formulation, sellers have either high or low quality goods and can be honest or dishonest

in their advertisements. Seller behavior is only observed by the reputation system, which provides assessments of either seller honesty or type (quality of goods) to the buyers. They find that the reputation system increases both the probability that buyers find high quality goods and their equilibrium price. However, they conclude that the value of the reputation system is ambiguous: under certain conditions (namely, when there are too few sellers or the reputation system provides insufficient information) the buyers see a net decrease in welfare with the reputation system.

A common thread runs throughout all of these analyses: reputations become important under conditions of adverse selection and/or moral hazard. When one party to a transaction has only incomplete knowledge of the other side, inept or dishonest agents can more easily compete since they cannot be distinguished from honest and competent ones. Reputation helps to mitigate this effect by providing a prediction, albeit a noisy one, of a potential partner's future behavior.

While most of these works do not address the role of reputation systems in peer-to-peer markets, clearly this application has features in common with the specific models discussed above. Given anonymity on-line, there is clearly potential for adverse selection — without additional information, it is impossible to discriminate between a "good" and a "bad" agent. Reputation facilitates this kind of discrimination and therefore accrues positive value, which is also true of on-line reputation, as discussed in the next section. It is this positive value to a good reputation that we believe provides an incentive to cheat the reputation system, which is not discussed in these models.

2.1.2 Marketplaces

As can be expected, the literature of markets is orders of magnitude larger than that of reputation in the economics community. However, most of the mechanisms at play in marketplaces are only of tangential relevance to the function of the reputation system. Therefore, we briefly mention a few important theoretical results but concentrate on looking at empirical studies of actual user behavior on on-line market sites.

Theoretical Market Models

The classic paper by Akerlof (3) introduces the problem of adverse selection that arises due to information asymmetries present in the market. In his example of the used car market, there are two types of cars, reliable ones (the "cherries") and unreliable ones (the "lemons"). While a buyer would gladly pay more for a cherry than for a lemon, the two types are impossible to distinguish at the time of the sale, so all cars are sold for a single expected price that accounts for the distribution of car quality. In such a market, the seller of a cherry cannot get a fair price and is eventually inclined to leave the market, resulting in an environment where only lemons are available.

On-line markets exhibit exactly this same information asymmetry — arguably, to an even greater degree than in Akerlof's example. After all, one can always employ a expert mechanic to inspect a used car before purchase, but usually cannot even physically see goods purchased on-line until after the sale is complete. Akerlof briefly posits four possible counteracting institutions for managing this problem: guarantees, brand names, chain stores, and professional licensing. Of these, chain stores and brands are both directly related to reputation: brand names and chains allow firms to establish a reputation for quality (or its lack) permitting consumers to make more intelligent assessments instead of just assuming average quality. Guarantees play a role in on-line markets — many sellers accept returns of unsatisfactory goods — but anonymity and jurisdictional issues make such guarantees little more than cheap talk. Professional licensing has, to date, played no role in on-line markets, though one could imagine a system whereby market participants are bonded in the fashion of tradespeople by an organization that helps to intermediate in case of disputes.

While the pricing mechanism used by the market is generally of little concern to the reputation system, we would nevertheless prefer a mechanism that prices items efficiently. Vickrey (108) looks at the optimality of various pricing schemes and concludes that the ascending price English auction (or equivalently, the second price sealed bid auction) is Pareto optimal for auctions of only a single item or a collection of identical items when we assume that bidders have independent, private valuations of the item being auctioned. A similar analysis can be found in Milgrom (87), which goes into more mathematical detail concerning bidding strategies. Furthermore, English auctions require less information gathering by bidders than do descending price Dutch auctions. In the English

auction, the bidders need only determine their valuation of the object, which then becomes the optimal bid. In a Dutch, or descending price, auction, bidders must also know each other's valuation distributions, and determining the optimal bid quickly becomes intractable. Related problems arise when the bidders' valuations are not independent. Such a situation arises frequently in auctions for rare collectibles and other one-of-a-kind items. An individual bidder's valuation will be influenced by the items resale value, which is signaled by the other bidders' behavior.

The auctions used by eBay are modeled after the classical English auction. The use of finite bid increments may introduce slight deviations from theory — the equivalence between the English auction and the sealed bid second price auction only holds if the bid increment can be made arbitrarily small — however, bid increments are small (typically $1 or less) in practice. In classical English auctions, there is no set end time but bidders may not re-enter once they stop bidding. Auctions on eBay are for a fixed duration, with bidders permitted to make additional bids up until closing. The implications of these differences are investigated by (7). There is a great volume of research into specific features of on-line auctions and their similarities and differences to classical auctions. A survey of this work and its results can be found in (92).

We concentrate on on-line auctions primarily because reputations are most heavily used at auction sites. However, the choice of pricing mechanism is substantially orthogonal to the functioning of the reputation system. The pricing mechanism may enable specific bad behaviors, such as shill bidding (112) in auctions, not present with other mechanisms. However, no known pricing system can overcome the potential for adverse selection in on-line markets, so reputations will always be necessary. The only requirement for the pricing system is that it must allow users with high reputations to buy and sell more easily or at a better price than those with poor reputations. Readers interested in further information about on-line markets beyond auctions and how they differ from traditional markets are encouraged to begin with the survey in (37).

Empirical Studies

While theoretical examinations of marketplace and reputation dynamics can be illuminating, real markets exhibit a degree of complexity impossible to capture in a model. Because of the

massive quantity of publicly accessible information available on eBay, it has become a natural target for empirical studies of peer-to-peer markets. Most of these studies seek to determine fundamental properties of the market, such as the role of adverse selection and moral hazard, the impact of reputation on prices, and the influences on bidder behavior. In this section, we briefly discuss some of the results discovered by these studies.

Resnick and Zeckhauser (95) conducted an early, influential study of eBay using a large dataset consisting of all transactions from February 1, 1999 to June 30, 1999 along with all feedback as of June 30, 1999. They examine the frequency of interactions, the rate with which users leave feedback, and the effectiveness of the reputation system. In their study, sellers transact about 2.7 times more often than buyers. Sellers are slightly more likely (60.6%) than buyers (52.1%) to leave feedback for their partners. When feedback is left, it is almost always positive: buyers leave positive feedback 99.1% of the time and sellers 98.1% of the time. They also find some evidence of retaliatory negative feedback: buyers retaliate for a received negative or a neutral about 7% of the time, and never leave a positive after receiving a negative (most often they leave no second feedback). Sellers retaliate against 19% of buyers that have left them a neutral or negative. They find that buyers leave the first feedback twice as often as sellers, consistent with the theory that sellers defer feedback in order to wield the threat of retaliation.

In addition to summary statistics about user behavior, Resnick and Zeckhauser also examine the effectiveness of the reputation system. They find that reputations do predict future behavior, but that the prediction is rather noisy: a user could cut the rate of problematic transactions by nearly two thirds, but only by avoiding 50% of unproblematic transactions. In their study, both the seller's positive and negative feedback have a significant effect on whether or not an item will sell, but they do not study the impact of reputation on price.

Resnick and Zeckhauser conclude by offering several observations about their study. They posit that eBay and its reputation system function because it maintains a "High Courtesy Equilibrium," where participation in the reputation system is high and users are reluctant to level negative criticism against their partners. While such an equilibrium does explain why users leave feedback readily and are reluctant to leave negative feedback, it also appears very susceptible to attack by malicious users. Two phenomena explain how eBay is able to maintain the high courtesy equilibrium: paying

initialization dues and stoning bad behavior. Because a new user has no reputation (and may, in fact, be a malicious user that has discarded an old identity), new sellers must pay initiation dues, such as reduced sale prices, lower probability of sale, or spending some time as buyer, until they have accumulated a sufficiently high reputation. The cost of building this baseline reputation also helps to encourage reputation maintenance through honest behavior. Defectors are punished by "stoning," a phenomenon where users are more likely to leave negative feedback for someone who already has a few black marks than for someone whose reputation is otherwise perfect.

Because of the size and breadth of the dataset Resnick and Zeckhauser used, their report provides a unique insight into the functioning of eBay. While other studies investigate specific mechanism in more detail, they cannot provide as complete a picture of the entire market. We use many of their summary statistics to calibrate our marketplace simulator (see Appendix A) to model eBay as closely as possible.

Because most other studies have been forced to use data crawled from eBay, rather than a comprehensive dataset like in (95), they tend to concentrate on more specific characteristics of the market. Broadly speaking, these studies can be divided into to two types: the first, and more common, type of study uses some variety of hedonic regression to determine the factors that affect pricing, number of bidders, or other market mechanisms. To control for price variability across items, these studies typically look at either a very specific type of commonly sold item or a category where there are published "book values" for the items.

The second type of study conducts a controlled field experiment by buying or selling specific types of goods. These studies are able to better control for external variables than the regression studies. However, because conducting the experiment is more costly than scraping the eBay web site, sample sizes tend to be smaller.

Bajari and Hortaçsu (7) investigate the factors that contribute to bidding and pricing behavior on eBay. They look at auctions of collectible coins, using book values from a contemporaneous coin collecting magazine to control for pricing differences. In particular, they investigate the relative impact of common values and private values on bidding behavior, and conclude that there are elements of both models at play. Their model explains the often observed phenomenon of "snip-

ing," where bidders wait until the closing minutes of the auction before bidding, as a consequence of agents responding optimally to the common value component of the auctions. Reputation also has significant effects in this study. Both a seller's overall reputation and the amount of negative feedback affects the number of bids an auction receives, but only overall reputation has a significant effect on auction closing price in this study.

Dewan and Hsu (28) study auctions for stamps, looking for evidence of adverse selection. They also normalize prices by comparing them to book value, but further compare normalized prices on eBay to prices sold at auctions run by a well-known dealer with a reputation for fair and accurate assessments of stamp quality. They find that items at eBay sell for less than at the dealer's auctions, consistent with significant adverse selection. This adverse selection gets worse with higher value items. However, the differences in prices the sellers receive in both markets are not as dramatic, suggesting that the intermediary captures most of the value added by its more accurate quality assessments. Reputation, posited as a solution to the problem of adverse selection, does not have a large impact in this study. While statistically significant, the economic benefit to a seller of a good reputation is relatively small — sellers with reputations in the 90th percentile are able to command only a 11.5% premium over sellers with 10th percentile reputations.

Cabral and Hortaçsu (15) investigate the role of reputation in greater depth by looking at eBay auctions for four items: two types of collectible coins, a specific model of IBM laptop, and a type of "Beanie Babies" toy. Across all four items, they observed that the first negative feedback has the greatest impact. After receiving a first negative feedback, the growth rate of a seller's business reverses from +7% per week to -7%. Subsequent negative feedbacks have no significant effect. However, buyers leave negative feedback significantly more readily for sellers who already have a negative feedback than for ones with pristine reputation. They investigate reasons why buyers might be more willing to leave negative feedback for sellers that already have a tarnished reputation. Among the theories they tested is that sellers are more likely to retaliate for a first negative feedback. However, while the overall retaliation rate was high — about 40% — there was no significant difference between retaliation for the first and subsequent negative feedbacks. They end up concluding that these mechanisms do not explain the data and that the higher frequency of negative feedback reflects actual changes in seller behavior. Finally, they look at the influence of reputation

on seller exits, and find that the numbers of positive and negative feedbacks are respectively negatively and positively correlated with the probability that a seller exits the market. Their observations agree roughly with several popular models of markets with reputation. In particular, the significant and apparently irreversible change in seller behavior after the first negative feedback corresponds quite well with the theoretical predictions of behavior in a market with both adverse selection and moral hazard. (74) Their data also lend credence to the "initiation dues" and "stoning bad behavior" theories of (95).

Khopkar et al. (71) present data that both confirms and contradicts findings in (15). While they observe the same increase in the likelihood of subsequent negative feedback after a seller receives a first negative, they attribute this effect to changes in both seller and buyer behavior. They find that the typical cause of the initial negative is due to a temporary decline seller quality. However, they present both anecdotal and empirical evidence that buyers are more willing to "stone" a seller that has received a recent negative. During this period, sellers have two options: either they raise their quality to salvage their reputation or they let quality slip since they can no longer command as high of a price premium. Neither of these alternative emerged as a dominant strategy. As in (15), Khopkar et al. find that sellers are significantly more likely to leave ("self-select") after receiving a negative feedback.

Most empirical studies look at easily quantified metrics like the amount of positive and negative feedback or the aggregated reputation score. Ghosa et al. (49) take a novel approach and analyze the influence on item price of the text comments that accompany feedbacks. They first decompose reputation into many dimensions, such as "product," "shipping," and "service." They then use a text mining application to extract terms that indicate positive and negative scores on these axes. Not surprisingly, terms like "wonderful product" have a strong positive correlation with increased seller pricing power, while terms like "never responded" have the opposite effect. In addition to the insights into the market's functioning, this experiment also suggests possible methods for aggregating non-numeric reputation information.

Houser and Wooders (60) studied the effects of reputation on the sale prices of a specific type of Pentium III microprocessor and, like many previous studies, find that both positive and negative feedback has a statistically significant but economically small effect on sale prices. In contrast

to (7), they suggest that a private value model is more appropriate than a common value model for eBay auctions. This difference is likely due to the vastly different nature of the items studied — stamps are often purchased speculatively for possible future resale, while microprocessors are usually acquired for use.

Lewis (78) also presents results that differ significantly from other studies. He finds that seller reputation has very weak effect on final sale prices at eBay Motors, the used car segment of eBay. In both his hedonic regression study and in the market model he constructs, the amount of data the seller reveals about the car (e.g. photos, inspection reports, etc.) has the strongest effect on sale price. He argues that seller revelations of a car's quality allows buyers to discriminate among items in this otherwise classic example of a market with adverse selection. It is unclear, however, how broadly these results can be applied to general on-line markets because used cars are rather unique in the number of credible services available for assessing their quality.

The final regression-based study we examine, by Lucking-Reiley et al. (80), also uses collectible coin auctions to explore the dynamics of pricing on eBay. They find that reputation has a significant effect on the price a seller receives, and that negative feedback has greater influence than positive feedback. This result contradicts (7), which also studied coin auctions and finds that negative feedback did not have a significant effect on price. Lucking-Reiley et al. also find that minimum bids and reserve prices have a positive effect on final sale price and that prices are higher in longer auctions. This latter result also contradicts other studies that use auction length as a regressor (28; 60), which found either no significant effect or a slightly negative effect of longer auctions on final sale price.

Resnick et al. (97) conduct a field experiment investigating the role of reputation in on-line auction prices where variables are controlled by design rather than through statistics. They sold approximately 200 paired lots of vintage postcards, with one member of the pair sold by an established seller with a good reputation (more than 2000 positives and one negative) and the other sold by the same seller but using a newly constructed identity with no reputation. They find that, on average, buyers are willing to pay about 8.1% more when buying from the established identity. In a second phase, they added negative feedback to some of the reputations of the low-experience identities created during the first part of the experiment and measured the difference in performance

between sellers with and without negative feedback. In this experiment, they found no significant difference between sellers with and without negative feedback.

In the experiment of Jin and Kato (62), researchers bought 107 collectible baseball cards from various sellers on eBay and had the received cards professional appraised. Unlike the used car auctions studied in (78), much of the seller revelations in the baseball card market is self-described and difficult to authenticate. Because of the low credibility of sellers' descriptions, they find significant evidence of fraud, or at least inflated claims. They find no significant differences in appraised quality between cards whose descriptions put them among the best-ranked ones available for sale and those whose advertised quality was around the median. However, the cards with the highest claimed ranks (9–10) command a 23.4–53.9% price premium over ones with lower claims of quality. In addition, 9 of the 11 fraudulent transactions they observed were for cards with the highest seller assessed quality. Regarding reputation, they find that higher reputations are negatively correlated with fraud and default but not with item quality. Their experiment helps to explain patterns in other studies, where a better reputation has a greater effect on the probability of sale than it does on the final selling price.

On-line reviews have many characteristics in common with reputations — one can think of a collection of reviews more or less as an item's reputation. Chevalier and Mayzlin (22) study the impact of reviews on the sales rank of books at Amazon.com and barnesandnoble.com, two leading Internet booksellers. They find that both the number and type of reviews had significant effects on sales at the two stores. The effects are larger and more significant at Amazon.com, which the authors attribute to the larger size and greater vibrancy of its user reviewer community.

While a very interesting window into the functioning of on-line markets, these empirical studies raise as many questions as they answer. Not surprisingly, none find real markets to work exactly as predicted by the theoretical models discussed above. In fact, different studies find evidence supporting different and often contradictory models. While some studies (7; 78) suggest that a common values auction model is appropriate, others (60) support the private values theory. Similarly, reputation seems to play a slightly different role in each study. Most likely, these discrepancies are

due to the widely varying data set sizes, choice of items studied, and the time frame of the study[1]. However, we do observe several broad themes running through these works:

Adverse selection and moral hazard play a role at eBay. While the magnitude of their effect is difficult to quantify, the difficulty of providing credible signals of item and seller quality make it possible for sellers to commit fraud or to turn eBay into the classic "market for lemons."

Reputations are effective, but far from perfect. In these studies, reputations were consistently significant to some extent, although their economic effects were quite variable. Most studies found some monetary value to a good reputation, and both buyers and sellers seem to care about reputations.

eBay is far from homogeneous. While there is certainly commonality in user behavior across these studies, the differences are can be striking. While eBay buyer bidding behavior exhibits signs of both common value and private value strategies, the relative balance between these two approaches shifts depending on the item being auctioned.

Several additional studies we do not cite are summarized in (27), which also clearly presents many of the challenges and opportunities facing reputation systems in peer-to-peer markets.

Fraud and Reputation Systems

While many of the above studies address how reputation systems can reduce the severity of fraud and deception in the market, the works in this section deal with the ways people try to game the reputation system itself.

As mentioned in Chapter 1, a major issue confronting peer-to-peer markets is the fluidity of identity on-line. Friedman and Resnick (45) investigate this problem and propose some solutions. They outline a cryptographic protocol that permits persistent pseudonyms: on-line identities that maintain privacy and anonymity yet providing a strong guarantee of one-to-one correspondence to off-line individuals. Unfortunately, in the time that has passed since this proposal, there does not

[1]The data in these studies were collected at various times from 1998–2005. During that time period, eBay made many changes to their site design, auction options, fees, and reputation system that may affect our ability to compare one study's results to the next, even if the item types are similar.

appear to be any movement towards making this type of service a reality, despite its obvious benefits for many on-line applications, including peer-to-peer markets. Since robust identity guarantees are not likely to be implemented in the near future, existing systems must simply adapt to a reality where identity is unverifiable.

Bhattacharjee and Goel (9) derive sufficient conditions for a purely economic means of preventing sybil and collusion attacks in eBay-like reputation systems. They show that if transaction fees are set high enough, it is impossible to profit by increasing one's reputation through a "ballot stuffing" attack using sybils or a network of colluding users. However, it is not clear what effect these fees might have on overall participation in the market. If high enough in absolute terms, users may simply prefer the alternative of lower costs but higher risks. Also, setting the optimal value for these fees requires an assessment of the price premium that a good reputation provides. As we discuss in the previous section, there is little consensus concerning the degree to which reputation affects pricing. Most likely, reputation's impact will differ depending on the type of good being sold and the individual sub-community within the larger market, making an optimal fee structure very hard to determine in practice.

Finally, Calkins (16) provides a legal analysis of eBay's reputation system. In her study, she finds that feedback forum reputations systematically overstate the safety of eBay's market. There is also evidence that this overstatement is deliberate, or at least deliberately disregarded, because it helps eBay attract new members by making the market appear safe. Resnick et al. (95) observe a similar contrived cheerfulness throughout the eBay site, but attribute it to the more admirable goal of fostering the high courtesy equilibrium needed for the reputation system to function. Calkins also argues that current regulations force eBay to play a fairly passive role in policing their market, because to do otherwise would open them up to considerable liability. While less technical than many of the above studies, this study presents a very sobering picture of the reality of eBay and describes the important external forces that are difficult to describe in a purely econometric analysis.

2.2 Computational Approaches to Reputation

While reputation has a long history of study in the economics community, it is a relatively new field among computer scientists. Prior to the rise of large-scale networks, most of the focus was on "hard" security: technologies like cryptography, access control, and security policies for keeping machines and data safe from untrusted users. The question of *who* to trust was not especially interesting. A user or administrator would make a list of trusted people and the level of access they should have, then use the list to implement a security policy for a resource.

In the connected would brought to us by the Internet, such approaches simply do not scale. It is ridiculous to suppose that anyone could enumerate the level of trust a particular user should have towards every other user on eBay, allowing a simple security policy to be coded. We need so-called "soft" security: a mechanism (such as a reputation system) for evaluating the trustworthiness of previously unknown others.

2.2.1 Conceptions of Trust and Reputation

Much of the early work in this field concerns itself with defining exactly what is meant by "trust" and "reputation" in the context of distributed applications. Marsh (84) attempts to distill much of the results about trust and reputation from the economics, sociology, and psychology communities into a concept of trust that can be applied to distributed multi-agent systems. In his formulation, trust ranges from +1 (complete or "blind" trust) to -1 (complete distrust). He discusses how trust estimates are built up through experiences colored by an agent's disposition, the implications of temporal effects like changing behavior and agent memory, and the principles that guide reasoning about trust by rational agents. He demonstrates the utility of his formalism by using it in an iterated prisoners' dilemma simulation.

One shortcoming of Marsh's treatment of trust is that it is entirely experiential. However, in large-scale applications, evaluating the trustworthiness of strangers is perhaps more important than a fully rational treatment of how we form our own trust judgments based on personal experience. Reagle (66) defines reputation as "the amount of trust an agent has created for himself through interactions with other agents." In other words, an agent's behavior in the marketplace is one of the

factors influencing the trust other agents have in it. Reputation systems collect this information and label the aggregation the agent's reputation. Like we saw in the economics literature, Reagle treats reputation as an asset: something that assists in the creation of value for an agent.

Reputations can be built up through honest behavior and the "dues paying" mechanisms of (95), maintained through a consistent level of honesty, and spent by "trading on one's reputation" and using it to cheat. Dai and Finney (25) characterize a reputation in terms of the values these uses yield:

Operating Value the present value of the increase in future profits due to the agent's current reputation

Throw-away Value the profit an agent could make by throwing away his reputation and cheating all his customers

Replacement Cost the expected cost of recreating an equivalent reputation if the agent discards his current reputation

The actual value of a reputation will be the greatest of these three values. For a system of reputation to be stable, the operating value should be greater than the difference between the throw-away value and the replacement cost, otherwise the most rational course of action will be for agents to cheat as often as they can and then later replace their reputation in order to start the cycle over again. These values again suggest that in applications where identity creation is costless, new users should receive the minimum reputation. If a new user starts with a positive reputation in such a setting, then the throw away value is positive but the replacement cost is zero, meaning that continually creating new identities just to cheat can be profitable, although suboptimal if the operating value is higher.

Khare and Rifkin (70) adopt some of these principles to more traditional security concerns on the web. They point out that the volatile nature of identity of users and machines on-line prohibits the application of traditional security policies. When deciding who to trust, they suggest that one can use a policy that discriminates by principle (identity), by object (e.g. possession of a secret code), or by capability. Secure systems can be built on any of these three policy foundations; the choice should be the one that makes implementation easiest and most flexible. In the context of

peer-to-peer markets, clearly capability (as expressed through the reputation system) is the only possible choice.

2.2.2 Reputation Systems for Peer-to-peer Markets

Barber et al. (8) present positions on a number of important issues regarding reputation systems for distributed systems. They identify three challenges for reputation system research: determining what elements of agent behavior contribute to trustworthiness, building reputation systems that do not rely on interaction, and establishing benchmarks for evaluating reputation systems. While not demonstration of either theory or technology, they address many important issues, such as whether agent intent should play a role in determine trustworthiness. Many of the concerns they identify, notably the lack of benchmarks for this field, are still relevant today.

Resnick et al. (96) also outline many of the difficulties in constructing reputation systems for on-line markets. They look at several popular on-line systems like eBay and Bizrate.com and consider how these systems restore the "shadow of the future," a concept introduced by Axelrod (6) where cooperation emerges because agents are concerned about the future impact of their actions in the present. They identify the three major responsibilities of the reputation system: the collection, aggregation, and distribution of trust information and discuss the strengths and weaknesses of existing systems in each of these phases.

Dellarocas (26) proposes a simple reputation system for on-line markets and discusses ways to immunize such a system against attacks such as "bad-mouthing" and "ballot stuffing," where a group of users colludes to lower or raise a target's reputation through inaccurate ratings. While some of his suggestions, such as using more robust statistics than the mean during feedback aggregation, are very plausible, others, like controlling anonymity on-line, clearly are not.

Kasbah (19; 118), an experimental testbed electronic marketplace that predates eBay, consists of a community of intelligent agents that search for products, set prices, and evaluate trust on behalf of human users. To facilitate trust, Kasbah relies on two related reputation mechanisms: Histos and Sporas. Sporas is a simple feedback-based scheme that assigns global reputations based on user ratings. The Sporas aggregation algorithm results in reputation scores between 0 and 3000 with

a single rating having less effect on higher reputations. The Histos algorithm exploits the small-world structure of social networks formed in the Kasbah marketplace along with a transitive trust relationship to give users personalized trust evaluations. Agents in the Kasbah marketplace can choose to use one or both of these systems.

Yu and Singh (117) describe an early, sophisticated reputation system for electronic communities (including markets). They create a trust model based on the one proposed by Marsh (84), but with several important differences. Their system makes explicit the "initiation dues" and "stoning bad behavior" of Resnick and Zeckhauser (95) by designing their aggregation function such that reputations are difficult to build up but easy to tear down. They also describe a model for propagating trust information, where other users' ratings are weighted according to one's trust in the raters during aggregation. In addition to the system's design, they provide an evaluation of its performance in a well-designed simulator.

The Pinocchio system of Fernandes et al. (43) and their followup system, Jiminy (73) uses incentives to encourage participation in a reputation system. Before users can query the system for reputation information, they must first earn credit by rating their past interactions. For each rating, a user receives credit that can be traded for use of the system's aggregated results. To prevent users from abusing the incentive system by leaving random or inaccurate feedback, they use robust statistical techniques to identify suspicious behavior and penalize the perpetrators appropriately.

The Beta reputation system (64), replaces the simple averaging of eBay's percent positive feedback with a Bayesian update using a beta distributed prior. The Beta system also includes a scheme for discounting feedback according to the level of trust one has in the rater and a mechanism for discounting feedback over time. While based on solid principles, the creators of Beta present no evaluation under real or simulated market conditions, so it is difficult to gauge its effectiveness. However, the beta distribution as a prior for reputation data proves to be natural choice, and we use a variation of this type of Bayesian update in our EM-Trust and Bayesian PPF system described in Chapter 4. A similar scheme also plays a role in the CONFIDANT (13; 14) reputation system for ad-hoc networks.

2.2.3 Other Reputation System Results

While our focus is on reputation systems for peer-to-peer marketplaces, systems proposed for other applications also have relevance to the problems we face. In this section we examine several of these systems as well as some general results that influence the results in the remainder of this thesis.

The EigenTrust (68) system attempts to solve the problem of collusion among users of peer-to-peer file sharing networks. In a collusion attack, a group of users agree to leave each other positive feedback or to band together to leave other users negative feedback, with the goal of unfairly increasing or decreasing the targets' reputations. EigenTrust is essentially a straightforward application of the PageRank (93) algorithm to the trust graph formed by positive feedback ratings along with a novel system for computing PageRank scores in a fully distributed fashion. However, the evaluation of EigenTrust in (68) is deeply flawed and most of the claims of collusion resistance fail under close scrutiny. Nevertheless, PageRank does have some useful properties as a reputation metric, which we exploit in the design of the Relative Rank and RAW algorithms of Chapter 5. While these algorithms demonstrate resistance against sybil attacks, the more general problem of collusion, which EigenTrust mistakenly claims to have solved, appears to be far more difficult.

While not a reputation system per se, Feldman et al. (41) evaluate a number of techniques for encouraging cooperation in peer-to-peer file sharing networks. Their evaluation technique involves a use of the iterative prisoners' dilemma, which also works well as a model for marketplaces, so we believe their results to be more general than just the application domain they study[2]. In particular, they demonstrate that computing the maximum flow in the trust graph yields a sybil-proof reputation metric. Unfortunately, the best known max-flow algorithms are still far to costly to apply to realistically sized markets.

Cheng and Friedman (21) look at the theoretical aspects of sybil resistance. Their main result is that a necessary condition for a sybil-proof reputation algorithm is that it be "asymmetric." An asymmetric reputation algorithm is one where a user's reputation is evaluated differently depending on the user's relationship with the person querying his reputation.

[2]Some aspects of Feldman et al.'s model, namely that only positive behavior can be observed, does not apply to markets, but many of their results do not have strong dependences on this assumption.

2.3 Summary

Peer-to-peer marketplace reputation systems are built on a wide foundation of work stretching back several decades. In the economics literature, reputations are proposed as a solution to the problem of reasoning about optimal behavior under uncertainty, as in the chain store game, the iterated prisoners dilemma, and other similar settings. In theoretical studies of markets, reputations help to counteract information asymmetries in markets with adverse selection or moral hazard, preventing the decay into Akerlof's "market for lemons."

eBay's market provides an ideal laboratory for testing these many theories. In the last ten years, dozens of empirical studies that attempt to explain pricing, bidding behavior, and the role of reputations have appeared. However, what eBay giveth, eBay taketh away: the vast quantity of data that permits these studies also leads to many different, often contradictory conclusions. eBay is not a single homogeneous market, but rather a federation of smaller communities that are both connected to the whole and largely independent.

With the rise of worldwide public networks and the large scale applications they engendered, the computer science community has also embraced reputations as a means of fostering trust between strangers. Much of the early work in this field revolved around turning the economic and sociological notions of trust into concrete formulations that can be gathered, aggregated, and disseminated efficiently. Studies such as (95) and (62) have shown that reputation systems can be effective at preventing some types of on-line fraud, but clearly more work is needed.

As shown in most of the empirical studies, reputation matters: buyers pay attention to it, sellers worry about it, and a good reputation has measurable impact on bidder entry and pricing. Unfortunately, it is precisely the value of a good reputation that creates incentives for users to game the reputation system in order to inflate their own reputation or mark down that of their competitors. Several systems have been proposed for combating certain types of reputation fraud, namely sybil/collusion attacks (68; 41) and dishonest feedback. (43; 73) However, making reputation systems robust against attackers is still a wide open topic.

Chapter 3

The Effect of Retaliation

3.1 Introduction

Reputation mechanisms developed for a number of different domains tend to share common features. These systems aggregate untrusted information — usually user feedback — to help users make trust decisions. However, much of the evidence supporting reputation systems for peer-to-peer markets is anecdotal. Large markets such as eBay and Amazon.com Marketplace deploy reputation systems to foster trust among their users, but we must not ignore the fact that the actual forces at work building trust in these markets are complex and extend far beyond just the reputation system.

We are interested in answering several fundamental questions: in the absence of external forces, is a reputation system sufficient for encouraging a pool of self-interested agents to cooperate? If so, under what conditions does stable cooperation arise? How well or poorly do existing reputation systems meet these requirements?

To this end, we model the trading and feedback process as a series of simple games. We then use this model in an evolutionary simulation to conduct several experiments that explore the influence of reputation system characteristics on optimal agent behavior. The results of these experiment show that an ideal reputation system is, in fact, enough to build trust in a peer-to-peer market; however, commonly used reputation systems have flaws that severely restrict their ability to function successfully.

We demonstrate that user apathy and manipulation of the reputation system can have large effects on its performance. In an ideal world, a reputation's value would purely incentivize good behavior: to create or maintain a good reputation, a user would engage in honest interactions with his or her peers. However, since reputations are valuable to their holders, there are also incentives for users to exploit loopholes in the system in order to unfairly inflate their reputations. Even when such manipulation is difficult or impossible, systems typically lack incentives to leave feedback, so reputation accuracy suffers if too few users participate in the reputation process.

One reputation manipulation strategy we examine in detail is retaliatory negative feedback, which is commonly observed in the Feedback Forum of eBay. In the typical retaliation scenario, two users interact and one cooperates while the other defects. As encouraged by the Feedback Forum instructions, the victim leaves a negative feedback for the defector, damaging his or her reputation as a warning to other users. The defector then retaliates by leaving negative feedback for the victim, who did nothing wrong aside from complain.

One might expect that the effect of retaliation would be more negative feedback than in a system without retaliation. However, the observed results are more subtle. Because a single negative has a much greater effect on the Percent Positive Feedback (PPF) score of a low volume user, high volume users wield retaliation as a threat against their less-experienced partners. Low volume users do not want to risk their fragile reputations by leaving negative feedback, particularly since they receive no direct benefit from leaving feedback, so the net result is that poor behavior is systematically underreported.

In our model market, retaliation causes a dramatic reduction in the reputation system's performance. When agents are permitted to retaliate for negative feedback, the reputation system is no longer capable of ensuring cooperative behavior. Because there are other external forces encouraging cooperation in real markets, it may be difficult to naïvely generalize these results beyond this abstract model. Nevertheless, our experiments suggest that retaliation is more than a mere nuisance; rather, it seems to be a potentially dangerous flaw in existing reputation systems.

3.2 Marketplace Model

Our simulations model a simplified peer-to-peer market where agents are either buyers or sellers. For these experiments we do not include agents that both buy and sell, but such an extension should be straightforward. All agents trade in the same commodity.

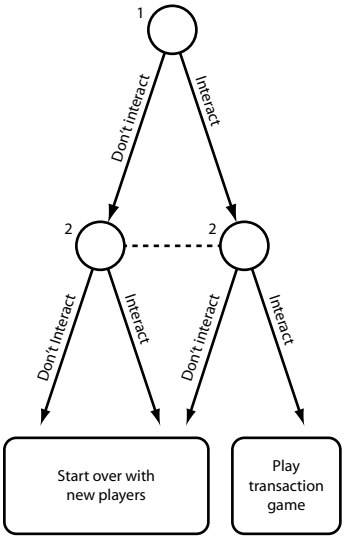
The trading and feedback process are modeled by the game shown graphically in Figure 3.1. This game consists of three separate periods: first the buyer and seller decide to interact, next they conduct a transaction, and last they participate in the reputation system. This trading game is played repeatedly by all the agents in the market.

3.2.1 The Interaction Game

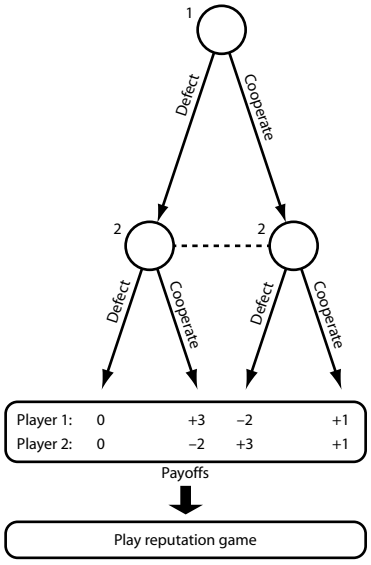
To begin the first period, we choose a seller that has an item to sell. The seller then plays the interaction game with each buyer that is waiting to purchase an item. The interaction game is a single stage, simultaneous move game where the two agents examine each others' reputations and decide whether they want to interact or not. If either agent chooses not to interact, the seller moves on to the next buyer and the buyer waits until another seller comes along. If a seller cannot find a buyer within a fixed period of time (one simulated "day" in our experiments), he or she gives up on the current attempted sale. Likewise, a buyer who cannot find a willing seller eventually gives up for the time being. If both buyer and seller agree to interact, they move to the next period and play the transaction game.

All agents make multiple attempts to interact periodically over the course of each simulation. The buy and sell rates are determined exogenously and are distributed according to a continuous Mandelbrot distribution (see Section A.4.1 in the appendix for more details).

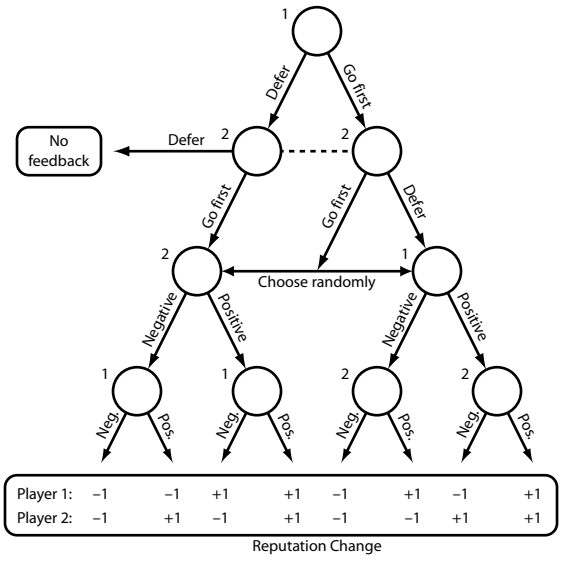
Most actions in the interaction game have no direct payoff. Choosing not to interact results in no penalty for either the refusing or the rejected party. Similarly, there is no reward for successfully finding a partner. Of course, the choice of partner will affect the outcome of the games played in the subsequent periods. We do give a small penalty (-10% of the reward for a successful transaction) to



(a) Interaction



(b) Transaction



(c) Reputation

Figure 3.1: The three games used in our model market. Circles represent moves with the player given by a number to the upper left. Arrows represent moves and dashed lines indicate information sets.

agents who give up because they cannot find a partner within one simulated day. This small penalty helps to promote interaction and prevents agents from refusing to participate in the market.

3.2.2 The Transaction Game

Once a buyer/seller pair have chosen to interact, they play the second period game, which models the trade transaction. Once again, this is a single stage, simultaneous move game that is quickly recognized as an instance of the classic Prisoners' Dilemma. Both players independently choose whether to cooperate (i.e. honestly complete their half of the transaction) or defect (attempt to defraud their partner). The payoffs are given in Table 3.1.

The Prisoners' Dilemma neatly models the process of trading with strangers in on-line markets. If both agents behave honestly, they each receive a modest reward (R). However, an agent that succumbs to the temptation to commit fraud receives an even greater reward (T) because he or she does not actually have to hand over payment or the item. The victim receives a sucker's reward (S) for giving up something of value while getting nothing in return. If both agents choose to be dishonest, then neither get anything (P). For the game to be an instance of the Prisoners' Dilemma, it must be the case that $T > R > P > S$. To realistically model actual trading, we add the additional constraint that $T \leq -S$, since it is unlikely that the scammer will receive greater value from the received item or payment than the victim lost. In our simulations, we use the values $T = 2$, $R = 1$, $P = 0$, $S = -3$.

The actual behavior of agents in the market are largely unaffected by the values for these payoffs, although there are some secondary effects. For example, larger values of T relative to R make agents somewhat more willing to try non-cooperative strategies while more a more negative S causes agents to be more cautious in the interaction game.

The Prisoners' Dilemma is a classic problem in game theory with a vast literature surrounding it. What makes it interesting is that even though there is an obvious strategy – always cooperate – that benefits both players, the dominant strategy for both players in the single iteration game is to always defect. The game thus has a single Nash equilibrium, where both players always defect, that yields a sub-optimal payoff (formally, the Nash equilibrium is not Pareto optimal). In an infinitely

		Player 2	
		cooperate	**defect**
Player 1	**cooperate**	R, R	S, T
	defect	T, S	P, P

Table 3.1: Graphical representation of the Prisoners' Dilemma that constitutes the transaction game. Each table entry gives the payoffs to player 1 and player 2 respectively.

repeated Prisoners' Dilemma game with two players, there are a subgame perfect equilibria that result in joint cooperation: for example, the "trigger strategy," where a player cooperates until the first defection and then defects ever after is one such strategy (see e.g. (50) pp. 88–96). Axelrod (6) performed several Prisoners' Dilemma tournaments and discovered that simple strategies like "tit for tat" (choose the same strategy as the other player did in the previous iteration) are most effective over the course of many iterations within a pool of players with different strategies, even though such strategies are not necessarily subgame perfect. Other studies have explored the role of forgiveness and reputation (91) among agents in similar iterative contexts.

One of the main goals of our experiments is to determine under what conditions a reputation system allows the market to converge to a cooperative state in the transaction game. The strategies employed by agents in such a cooperative market need not constitute a formal Nash equilibrium — tit for tat, for example, is not a subgame perfect strategy for the finitely repeated Prisoners' Dilemma. One of the fascinating aspects of the Prisoners' Dilemma is that purely rational (in a game theoretic sense) strategies are often sub-optimal, so players often choose to play sub-rational strategies that nevertheless result in better payoffs. Often, these agents engage in some form of trust building process to protect the delicate equilibria that result from these strategies. The analogy to trust and risk in on-line markets is obvious. We aim to explore whether or not reputation systems can effectively maintain this fragile cooperation over the long haul.

3.2.3 The Reputation Game

After the transaction finishes, the agents move into the final period, where they leave feedback for one another. The structure of this game depends on the choice of reputation system, unlike the

interaction and transaction games, which are constrained by the mechanisms of trading in an online market. We use a reputation game that is modeled after the Feedback Forum in use at eBay. In the first move, the players simultaneously decide whether or not they are willing to leave the first feedback. If both decide to leave the first feedback, we randomly (probability 0.5) choose one of the two players to go first. The first player then leaves feedback for the other player and can choose to leave either positive or negative feedback.[1] The first player must base his or her feedback decision on the players' behavior in the transaction and their reputation histories.

After the first player leaves feedback, the second player gets the opportunity to leave feedback. He or she can leave positive or negative feedback or choose to not leave feedback at all. In addition to the players' behavior and histories, the second player can also use the feedback received from the first player to decide what type of feedback to leave. It is this non-simultaneity of the reputation game that permits the existence of retaliation. A player that always leaves a negative for those that have given negative feedback hopes to cultivate a reputation as someone who retaliates and thus avoids future negative feedback.

The payoffs in the reputation game are indirect, which complicates the analysis of this game. Clearly, receiving a positive feedback makes it easier to win the next interaction game, while receiving negative feedback makes the interaction game more difficult. However, both of these payoffs will be a function of the players' interaction rates and current reputations — one feedback will have a large effect on the reputation of a new user, but very little effect on a player with a long history.

Leaving feedback (either positive or negative) has no direct payoff, but can have a dramatic effect on the performance of the market as a whole. Leaving accurate feedback — feedback that correctly indicates the recipient's transaction behavior — is of great value to the community at large. However, leaving retaliatory negative feedback may benefit the individual agent more, because it may dissuade future partners from leaving negative feedback.

Essentially, the reputation game results in another widely studied problem in economics and game theory: the tragedy of the commons. All players benefit from honest participation in the reputation system, but individual players can maximize their payoffs by leaving dishonest feedback

[1] We ignore neutral feedback because it is both infrequently seen and considered by most researchers to indicate a weak negative.

(e.g. retaliation) or by not leaving feedback. Unless something is done to prevent this selfish behavior, the "commons" typically collapses under the weight of all the freeloading. Aligning individual and social goals is often a goal of mechanism design; see the Pinocchio reputation system (43) for one proposed solution.

3.2.4 Theoretical Analysis

The combination of these three games results in a system that is difficult to analyze in closed form. However, we can look at simplifications of this model to gain some insight about its behavior. Consider the following simplified interaction and trading game:

- On each transaction, both players simultaneously choose an interaction threshold t_i and an honesty h_i.
- If $h_1 < t_2$ or $h_2 < t_1$, then the players do not interact.
- If $h_1 \geq t_2$ and $h2 \geq t_1$, the players interact by playing the Prisoners' Dilemma with the same payoffs as in the transaction game of Section 3.2.2. Player i cooperates with probability h_i.

The major change from the more complex model is that player honesty are immediately and perfectly observed, instead of being filtered by the reputation system.

Assuming that the payoff, P, when both players defect is zero, the expected payoff, π_i, for player i is:

$$\pi_1(h_i, t_i, h_j, h_j) = \begin{cases} (R - S - T)h_i h_j + S h_i + T h_j & \text{if } h_1 \geq t_1 \text{ and } h_2 \geq t_2 \\ 0 & \text{otherwise} \end{cases}$$

Subject to the assumptions of Section 3.2.2 that $P = O$, $T > R > P > S$, and $T \leq -S$, there are two Nash equilibria: $h_1^* = h_2^* = t_1^* = t_2^* = 1$ and $h_1^* = h_2^* = t_1^* = t_2^* = 0$. In other words, the players either always cooperate or always defect. Both equilibria are easy to verify: in the always defect case, if either agent cooperates with $h_i > 0$, then its expected payoff will decrease because $S < 0$. In the always cooperate state, both players receive a payoff of $R > 0$, but unilaterally playing an honesty $h_i < 1$ will prevent the player from interacting yielding a payoff of 0. The

interaction thresholds can vary from their equilibrium values without affecting the payoffs, a result that has important consequences for the evolutionary stability of this game.

A Nash equilibrium strategy S^* is defined to be *evolutionarily stable* if:

$$\begin{aligned} \forall i, \pi_i(S_i^*, S_j^*) &\geq \pi_i(S_i, S_j^*) \\ \forall i, \pi_i(S_i, S_j^*) &> \pi_i(S_i, S_j) \end{aligned}$$

By this definition, the cooperative equilibrium of this game is *not* evolutionarily stable. While cooperating, players can select threshold values off the equilibrium strategy that are neutral with respect to the payoff. However, once one player chooses an off-equilibrium threshold value, the other player has the opportunity to increase its payoff by reducing its honesty, allowing a pool of cooperators to be infiltrated by defectors.

It has been suggested that a model with an evolutionarily stable cooperative state would be more appropriate for modeling peer-to-peer markets. For example, a game where each agent's probability of interacting is equal to the honesty of its partner would have an evolutionarily stable cooperative state so long as $R \geq T/2$.

However, we feel that the use of thresholds for interaction more accurately models the way actual users behave in real markets. Each user determines a level of risk with which he is comfortable and bases his decision to interact or not on that choice. If told that a seller is honest only half the time, it seems unlikely that 50% of buyers would choose to interact.

Furthermore, there is no indication that cooperation in real markets is stable. While most users are honest, the fact that dishonest behavior is profitable despite the efforts markets expend to stamp it out suggests that a model with evolutionarily stable cooperation may, in fact, be an oversimplification. Reputation systems for real markets may need to deal with the fact that they are fighting defectors on unequal terms, a realization which may influence their design. In our experiments described below, long lasting cooperation does arise, despite the lack of evolutionary stability in the interaction game, suggesting that even if real systems suffer from a similar deficiency, stable cooperation is nonetheless realizable.

3.3 Simulated Evolution

Due to both the complexity of the three period interaction/transaction/reputation game and the ill-defined payoffs during the reputation period, a closed form analysis of this model is elusive. However, it is straightforward to simulate a marketplace where agents play these games repeatedly. Furthermore, such a simulation may reveal strategies of interest that may not meet the requirements of formal Nash equilibria or evolutionary stability.

To explore how the choice of reputation system affects marketplace performance, we create a simulator that evolves strategies for the three games described above. The simulation proceeds as a series of generations. In each generation, a pool of agents repeatedly plays the interaction, transaction, and reputation games via a mechanism described in Section 3.3.1. The agents' behavior in the three games are controlled by a set of parameters discussed in Section 3.3.2. Finally, at the end of each generation, the agents' performance is evaluated and we create new agents through a process analogous to sexual reproduction, as shown in Section 3.3.3. Over the course of many generations, unsuccessful strategies tend to die out, while successful ones multiply and dominate the agent pool.

3.3.1 Simulator Mechanism

The framework of the evolutionary simulator is based on the non-evolutionary simulator described in Appendix A and used in other experiments in this thesis. We briefly describe its mechanism here but refer the interested reader to the appendix for more details.

Within each generation, the simulator performs a set number of transactions. The timing of each agent's attempts to buy or sell are governed by a Poisson process with the individual rate parameters distributed according to the continuous Mandelbrot distribution described in Section A.4.1 in the appendix. The simulator maintains priority queues for buyers and sellers sorted by buy and sell times respectively.

Buy and sell times are determined exogenously in this simulation because we are interested in simulating markets that have a wide range buy and sell rates, just like real peer-to-peer markets.

Furthermore, we are interested in determining whether certain strategies are more applicable to certain interaction rates. Therefore, we further divide buyers and sellers into "small" and "large" classes. The small classes consist of agents whose interaction rates are less than the median rate while the large class is made up of those with rates greater than the median. The median rate is 0.2 transaction/day for buyers and 0.5 transactions/day for sellers, roughly the rates observed by (95) on eBay. Each generation starts with 1000 small buyers, 1000 large buyers, 400 small sellers, and 400 large sellers.

The number and type of agents does not remain constant during the course of a generation. An agents with a reputation low enough that it has a better chance of interacting as a new user will discard its identity. Two out of three such agents respawn and return to the market with a new identity but with all other characteristics unchanged. The remainder leave permanently. In addition to agents returning with new identities, we also add completely new agents to the market at a growth rate of 0.25% per simulated day. New agents' types and parameters are determined randomly and distributed according to the empirical distribution of the types and parameters in the market.

Agent interaction is handled much like in the non-evolutionary simulator, with sellers and buyers taken off their respective priority queues until a pair willing to interact is found. However, in the evolutionary simulator, the decision to interact is governed by the interaction game rather than a fixed threshold as in the standard simulator of Appendix A.

Once we find a pair of agents willing to interact, they proceed to play the transaction game. The agents choose to cooperate or defect and then observe the outcome of the game. Next, they play the reputation game to leave each other feedback. Finally, the simulator generates new buy and sell times for the participants and places them back in the appropriate queue.

Each generation consists of 50,000 transactions. Because the simulation often evolves so that agents are unwilling to interact with new agents without an established reputation, we also perform a burn-in of 10,000 transactions before running the generation. During the burn-in period, agents always agree to interact regardless of reputation, which creates baseline reputations for the initial pool of agents. The results of burn-in transactions are discarded when computing agents' values.

Parameter	Game	Description	Initial Values	
			Distribution	Mean
NUI	Interaction	New User Interactivity	Uniform(0,1)	0.5
LUIT	Interaction	Low-experience User Interaction Threshold	Beta(4,2)	0.667
HUIT	Interaction	High-experience User Interaction Threshold	Beta(4,2)	0.667
HON	Transaction	Honesty	Beta(18,2)	0.9
FP	Reputation	First Positive	Beta(9,1)	0.9
FN	Reputation	First Negative	Beta(9,1)	0.9
RET	Reputation	Retaliation	Beta(1,9)	0.1

Table 3.2: Agent parameters used for simulated evolution and their initial distribution in our experiments.

3.3.2 Agents and Strategies

Each agent in our simulation plays a strategy determined by seven parameters, given in Table 3.2 and described in detail below. Each of the parameters takes values in the interval $[0, 1]$ and together they describe a mixed strategy drawn from a subset of the possible strategies for the three games described in Section 3.2.

This parameterization does not permit agents to play all possible strategies in the three games. Rather, this choice of parameters constrains the strategies to a set frequently observed in real markets in order to make evolution tractable.

Like related techniques in genetic programming, this simulation is a form of gradient ascent: it tries to greedily optimize strategies by making small changes in the parameters. Without some constraints, the simulation tends to wander around parameter space and never converges. Previous experiments that permitted more general strategies in the reputation game (a total of 24 parameters) have had just this problem, so we constrain evolution along paths of interest by limiting the param-

eters in this fashion. The additional freedom gained by allowing more general strategies tends to have agents waste time playing obviously useless strategies that are never observed in real markets.

Interaction Parameters

The three interaction parameters (NUI, LUIT, HUIT) control how the agent plays the interactivity game. *New User Interactivity* (NUI) represents the probability that the agent will interact with a user that has no feedback history. In general, interacting with new users is risky, because new users are not necessarily all new: some will be agents kicked off the market for poor performance who are returning with new identities.

If the other player in the interaction game does have a reputation, the agent's strategy is controlled by either the *Low-experience User Interactivity Threshold* (LUIT) or *High-experience User Interactivity Threshold* (HUIT). Both represent the minimum reputation value that a potential partner must have before the agent will be willing to interact. LUIT controls behavior when the potential partner has received feedback from fewer than 10 partners, while HUIT is used when the potential partner has 10 or more feedbacks.

Transaction Parameter

A single parameter, *Honesty* (HON), controls the agent's strategy in the transaction game. As the name implies, this parameter indicates the probability that the agent will cooperate in the transaction game Prisoners' Dilemma.

Reputation Parameters

The three remaining parameters determine how the agent plays the reputation game. The *First Positive* (FP) and *First Negative* (FN) parameters indicate the probability that the agent will leave feedback first. FP is the probability of leaving a first positive when the agent's partner cooperated in the transaction game, while FN indicates the probability of leaving a negative for a defecting partner on the first move of the reputation game.

Description	Utility
Both cooperate	1.0
Both defect	0.0
Agent cooperates, partner defects	-3.0
Agent defects, partner cooperates	2.0
Unable to find a partner	-0.1

Table 3.3: Utility values for transaction outcomes.

The *Retaliation* (RET) parameter is the probability that the agent will retaliate for a received negative. When retaliating, the agent will leave a negative if it received a negative from its partner. If not retaliating, or if it did not receive a negative, then the agent will give feedback appropriate to the partner's behavior during the transaction game.

3.3.3 Evolution and the Value Function

At the end of each generation, we look at the transaction history of each agent and measure how well it competed in the market. Based on this metric, we create a new pool of agents for the next generation by keeping the most successful agents and creating new agents through the breeding process.

We compute each agent's value using the function:

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}}{N}$$

where v_i is the value of agent i, u_{ij} is the utility agent i gained or lost on its j^{th} transaction, and N is number of transactions that i attempted in the current generation. The value of u_{ij} is determined by the payoffs of the interaction and transaction games. The values we use in this experiment are given in Table 3.3. We normalize the sum of transaction utilities so that we can compare the success of agents that have different buy/sell rates.

Once we have computed the agents' values, we rank them and determine the pool of agents for the next generation. 20% of the new generation's pool of each agent class consists of the top ranked agents from the previous generation. The remaining 80% are new agents created through the

breeding process: we choose pairs of agents to be parents and combine their parameters to create child agents.

Under normal circumstances, we choose parents randomly from the previous generation's agents that received a positive value. We discard agents with negative values because their strategies are unprofitable, thus should not be passed on to subsequent generations. The probability of choosing an agent as parent is proportional to its value, so the higher ranked agents tend to have more children in the next generation than do low ranked agents.

In some generations, there are too few agents with positive values to form an adequately large breeding pool. In this case, we use the top 10% or top 10 agents, whichever is larger. In certain rare cases, fewer than 10 agents in a class survive to the end of the generation. Under such circumstances, we are forced to augment the breeding pool with randomly generated parents.

The breeding process itself is very simple: for each parameter in the child, we pick one of the parents randomly with probability 0.5 and copy its parameter value to the child. In 1% of these copy operations a mutation occurs and the child's parameter takes a random value.

3.4 Results

We perform two experiments with our simulator to demonstrate the reputation system's influence on the evolution of strategies in the market. Both experiments run for 10,000 generations or a total of 500 million transactions, disregarding burn-in transactions. We perform each experiment 20 times and average the results. The results of these experiments are summarized in Figures 3.2–3.7.

All experiments start with the simulator initialized with agents that represent a mostly honest, cooperative market. The specific initial distribution of agent parameters is shown in Table 3.2. We have several reasons for choosing to start the market in a cooperative state. In real markets, the community of initial users tends to cooperate. It is only over time that non-cooperative players infiltrate the market as they discover opportunities to profit at others' expense. Furthermore, it naturally seems harder to build a reputation system that can restore a non-cooperative market to cooperation than one that only protects an already cooperative market from takeover by defectors. While both

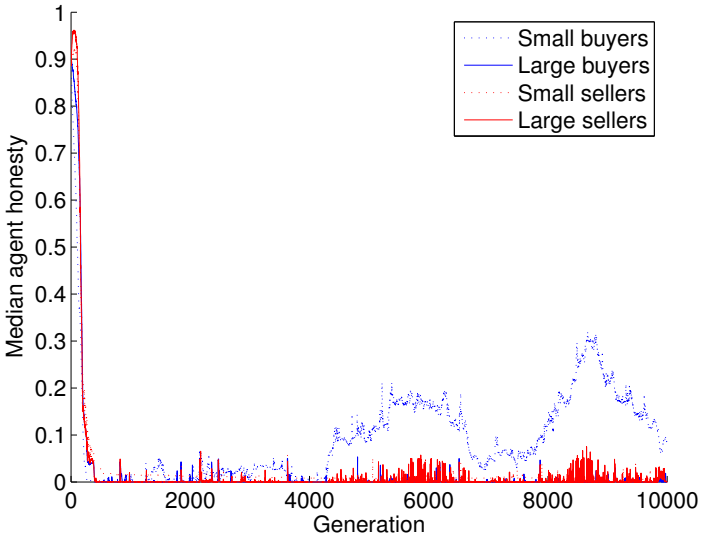


Figure 3.2: Evolution of agent honesty parameter with retaliation allowed.

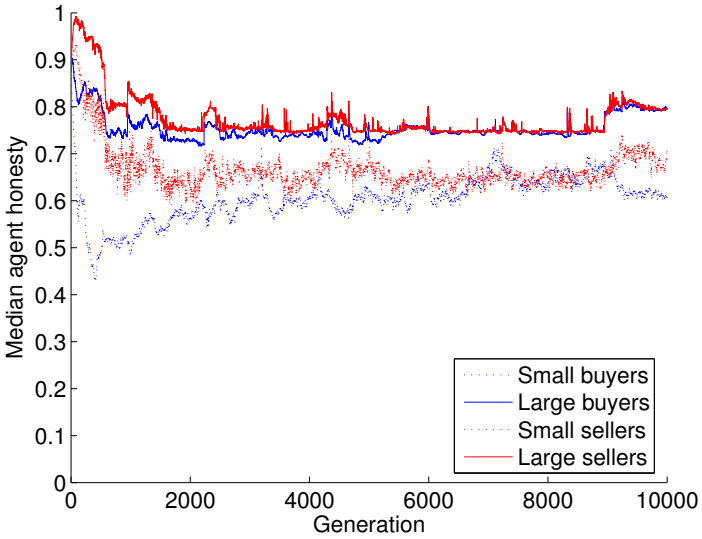


Figure 3.3: Evolution of agent honesty parameter with retaliation prohibited.

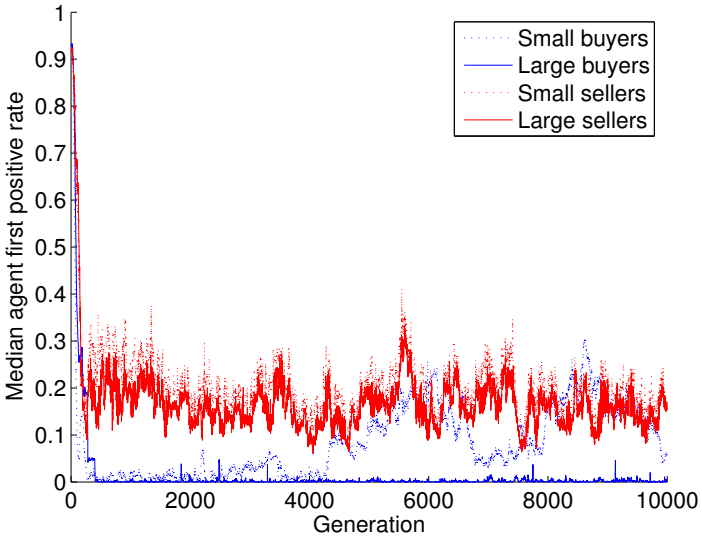


Figure 3.4: Evolution of agent first positive parameter with retaliation allowed.

Figure 3.5: Evolution of agent first positive parameter with retaliation prohibited.

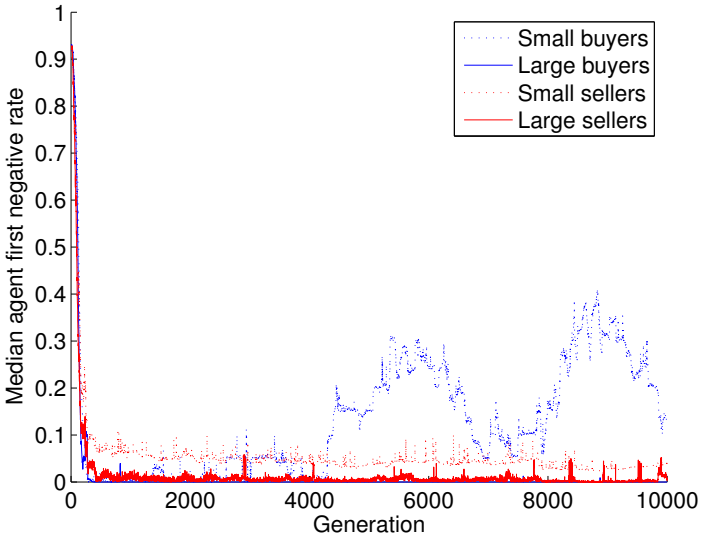


Figure 3.6: Evolution of agent first negative parameter with retaliation allowed.

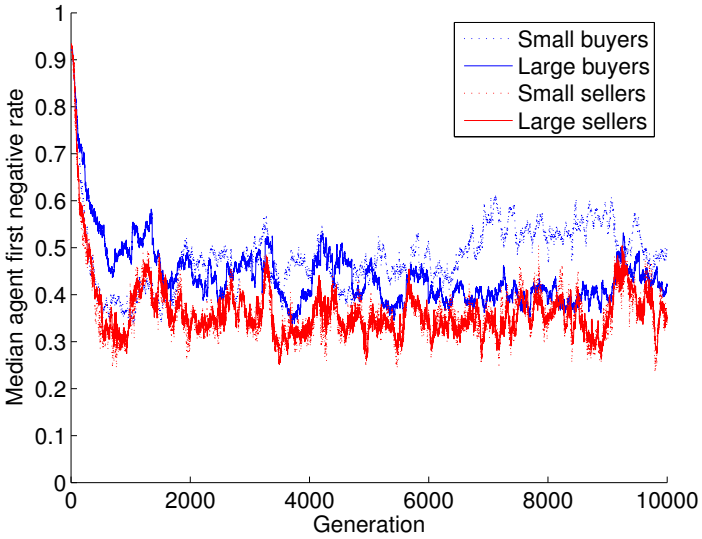


Figure 3.7: Evolution of agent first negative parameter with retaliation prohibited.

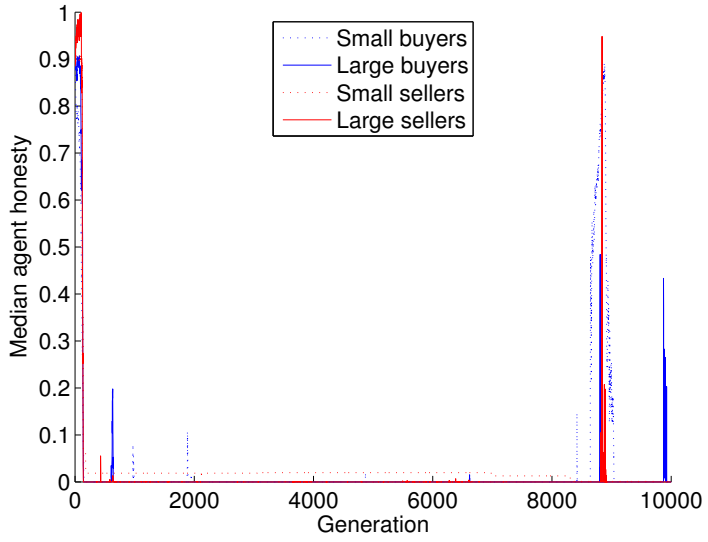


Figure 3.8: Evolution of median honesty in the first of two example markets with retaliation allowed.

are important, we feel that we must first understand the problem of maintaining cooperation before we can hope to make progress towards creating cooperation where none currently exists.

3.4.1 Performance with Retaliation

The first experiment shows the evolution of a market with retaliation allowed, using the model of Section 3.2. The results are shown graphically in Figures 3.2, 3.4, and 3.6.

The first graph (Figure 3.2) shows that the market rapidly converges to a non-cooperative state. By generation 500, all four classes of agents have a median honesty of less than 0.05. The short spikes are brief periods where one of the markets begins to move towards a cooperative equilibrium but is quickly dragged back down. The process of averaging multiple runs makes these spikes appear small, but individual markets actually occasional have spikes of fairly high honesty, as shown in Figures 3.8 and 3.9. However, these brief periods of cooperative behavior do not sustain themselves for longer than a handful of generations.

Participation in the reputation system is also poor in this market. Only the small buyers and sellers have a median first negative rate (Figure 3.6) significantly greater than zero, and even it is

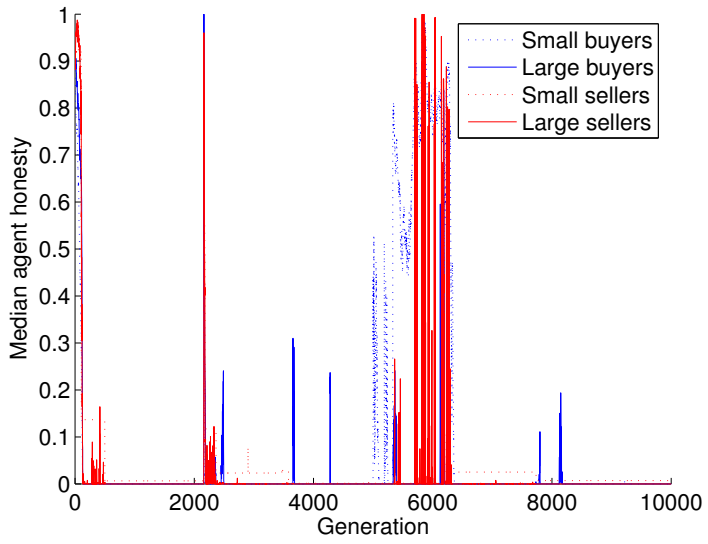


Figure 3.9: Evolution of median honesty in the second of two example markets with retaliation allowed.

very small. The first positive rate (Figure 3.4) is also quite low: near zero for buyers and only around 20% for sellers.

The markets in this experiment quickly evolve to have a moderate retaliation rate, shown in Figure 3.10. The retaliation rate has fairly high variance, which can be seen in the Figure 3.11, which shows how the retaliation rate evolves in a single representative market.

The results from this first experiment are intriguing but they raise as many questions as answers. Some of the evolved strategies, such as unwillingness to leave the first feedback, seem to correspond the observed behavior in the real world. Others, like the degeneration into a non-cooperative state, are at odds with real markets.

3.4.2 Evolution without Retaliation

One possible explanation for the results of the previous section is that the threat of retaliation causes the agents' hesitation to leave the first feedback. The lack of feedback makes it profitable to defect, leading to the collapse of the market. Similar effects have been observed in real markets, (95; 16) though the results are not as dramatic.

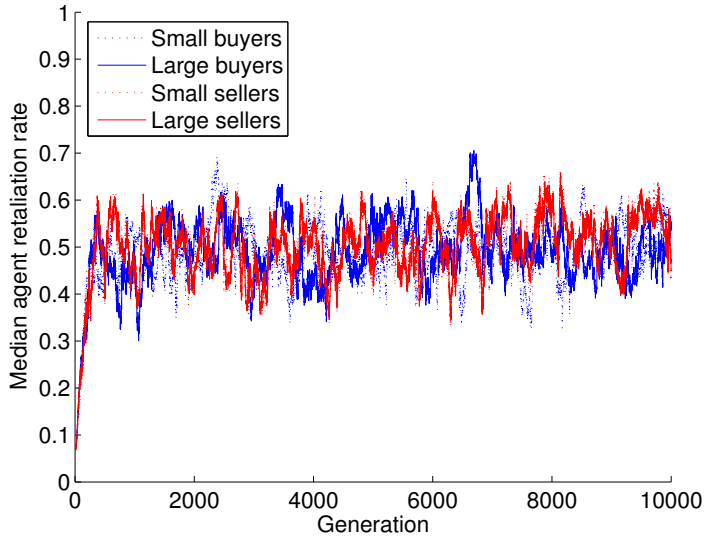


Figure 3.10: Evolution of median retaliation rate (mean of 20 runs).

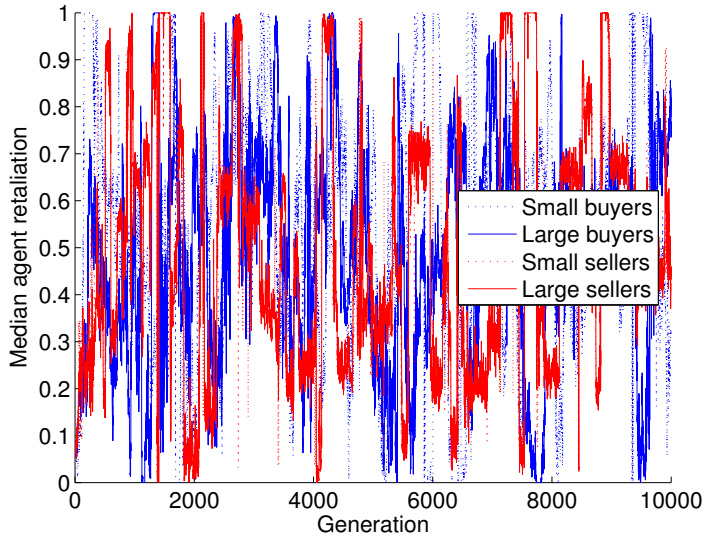


Figure 3.11: Evolution of median retaliation rate (single representative run).

To test this hypothesis, we stretch our evolution analogy even further and borrow a technique from biology: gene knockout. (17; 56) By fixing some of the agents' parameters to a known value and observing how the market evolves differently than the one seen in the previous section, we can infer the role the fixed parameters play. For this experiment, we fixed all agents' RET parameter to zero. Simply stated, we prohibit all agents from retaliating. The results of this experiment are shown in Figures 3.3, 3.5, and 3.7.

The markets that evolve when retaliation is prohibited are strikingly different from otherwise identical markets that allow retaliation. Most notably, the non-retaliating markets usually remain in a cooperative state. However, even without retaliation, cooperation is not assured. Of the 20 simulations, 14 quickly evolve to a cooperate state and remain there for the entire 10,000 generation duration of the experiment. Two markets oscillate: they first become cooperative, then turn non-cooperative, and finally return to cooperation. In the remaining four runs, the market evolves into a non-cooperative state and remains there for the duration of the experiment.

Stable Cooperative Markets

Figures 3.12–3.17 present the evolution of agent parameters in a representative market from among the 14 that evolved stable cooperation. In these markets, the large buyers and sellers become almost perfectly cooperative. The lower interaction rate of the small agents allow them to engage in a small amount of non-cooperative behavior yet still succeed in the market.

The participation in the reputation system is also quite different than when retaliation is permitted. Buyers nearly always leave a first positive, while sellers almost never do. This is likely due to the asymmetry in interaction rates: there are fewer sellers but they interact 2.5 times more frequently than the buyers. Because they interact less frequently, buyers have more incentive to leave a first positive so that they receive feedback in return and thus build their reputation. The more frequently interacting sellers can afford to demur, particularly since the buyers are eager to leave a first positive.

All classes of agents leave a first negative approximately half the time. While notably higher

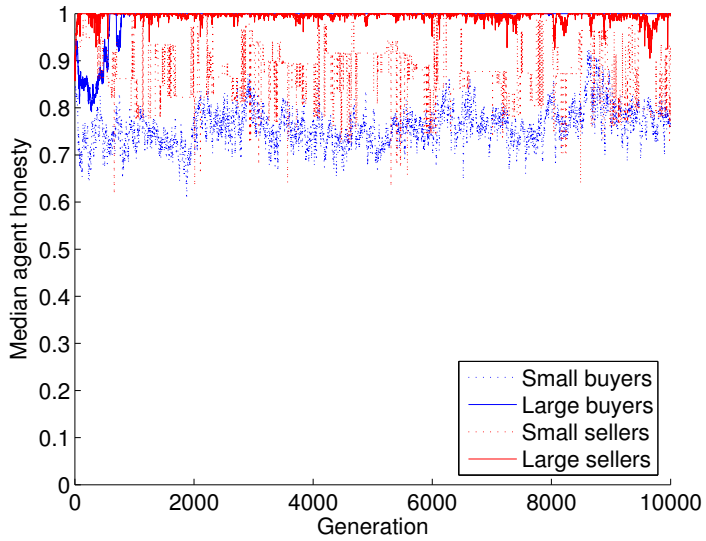


Figure 3.12: Evolution of the HON parameter in an example market without retaliation that evolves to a stable, cooperative state.

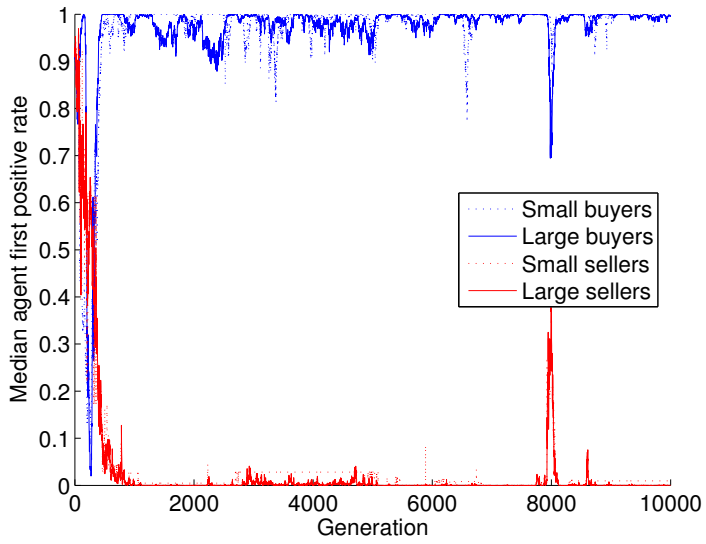


Figure 3.13: Evolution of the FP parameter in an example market without retaliation that evolves to a stable, cooperative state.

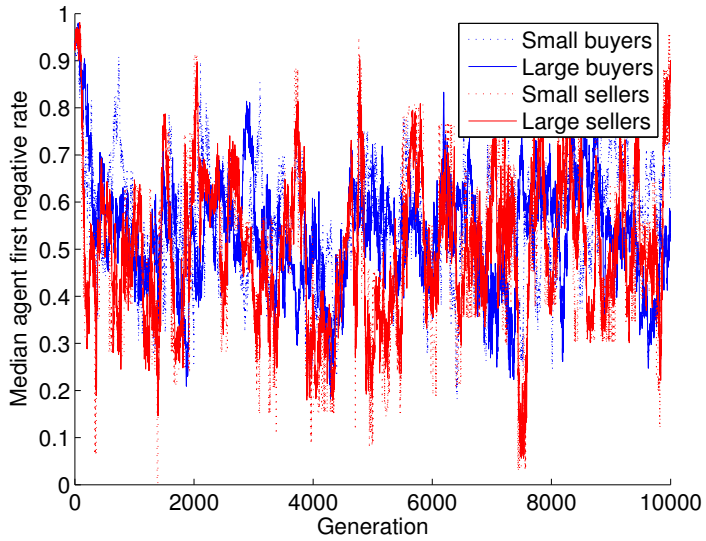


Figure 3.14: Evolution of the FN parameter in an example market without retaliation that evolves to a stable, cooperative state.

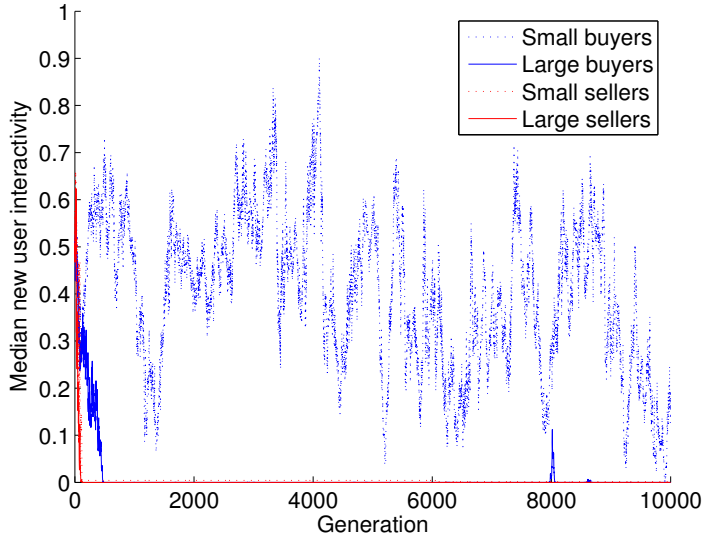


Figure 3.15: Evolution of the NUI parameter in an example market without retaliation that evolves to a stable, cooperative state.

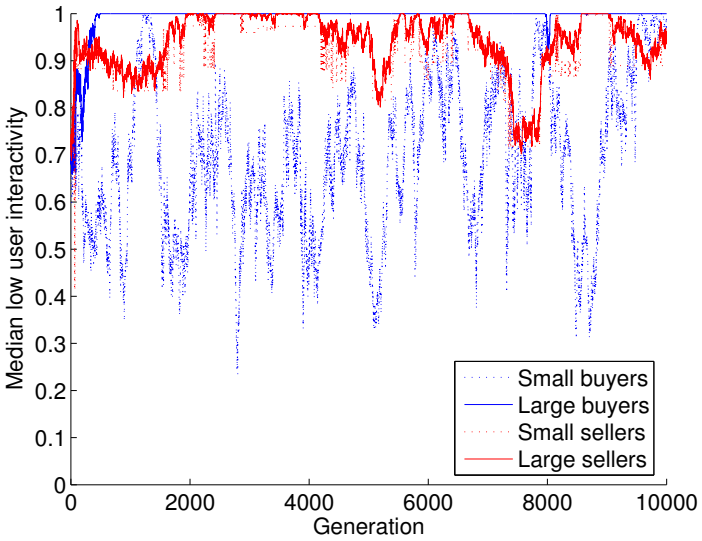


Figure 3.16: Evolution of the LUIT parameter in an example market without retaliation that evolves to a stable, cooperative state.

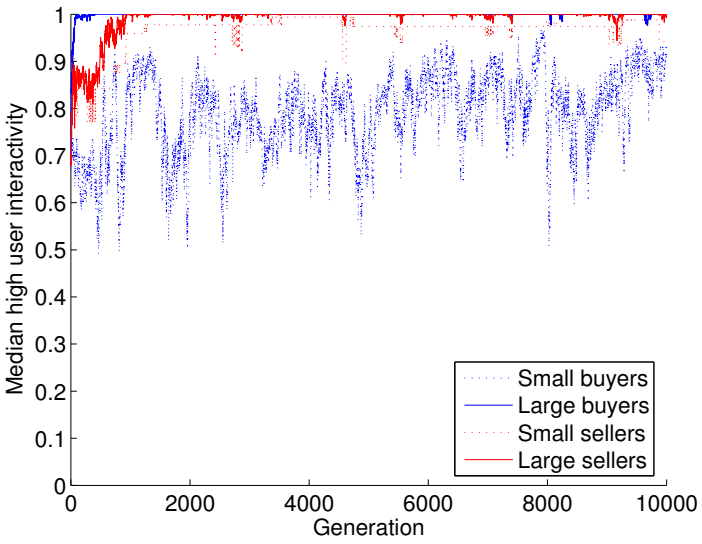


Figure 3.17: Evolution of the HUIT parameter in an example market without retaliation that evolves to a stable, cooperative state.

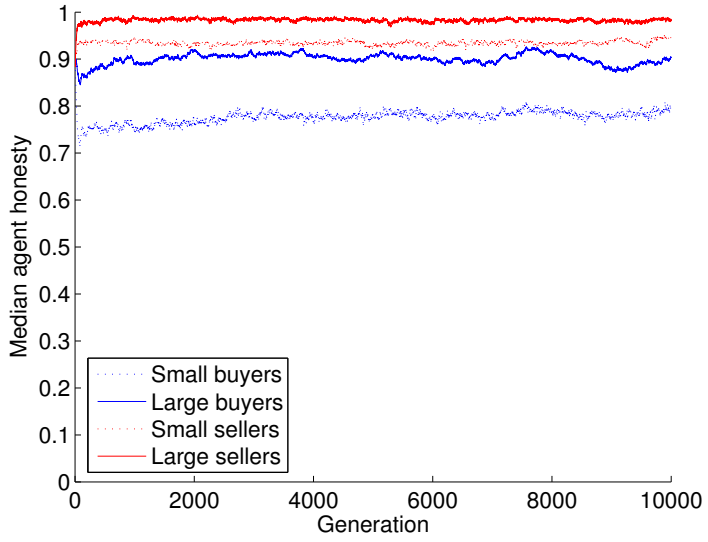


Figure 3.18: Evolution of the HON parameter in markets where users are forced to leave first negative feedbacks. Mean of 20 independent runs.

than when retaliation is allowed, it is not readily apparent why the rate is not higher. There is no clear penalty to leaving a first negative when there is no retaliation.

The only difference between the results of this section and the previous ones is that we prohibit retaliation. The complete mechanism for market failure when retaliation is allowed appears to be that retaliation chills negative feedback, which permits greater dishonesty. Dishonest agents have no incentive to leave positive feedback first (otherwise the retaliation threat has no teeth) so the reputation system completely collapses and so does the market.

To further verify this mechanism, we perform a second experiment where retaliation is allowed, but all agents are forced to leave a first negative feedback for defecting partners (see Figures 3.18 and 3.19). The results confirmed that negative feedback is one key element of a functioning reputation system. Qualitatively, the markets' evolution followed a similar course as when retaliation is prohibited, but participation in the feedback system is even better in this experiment: large sellers left first positives roughly half the time. Furthermore, all 20 markets are consistently cooperative in this experiment, but large buyer and seller cooperations rates are slightly lower.

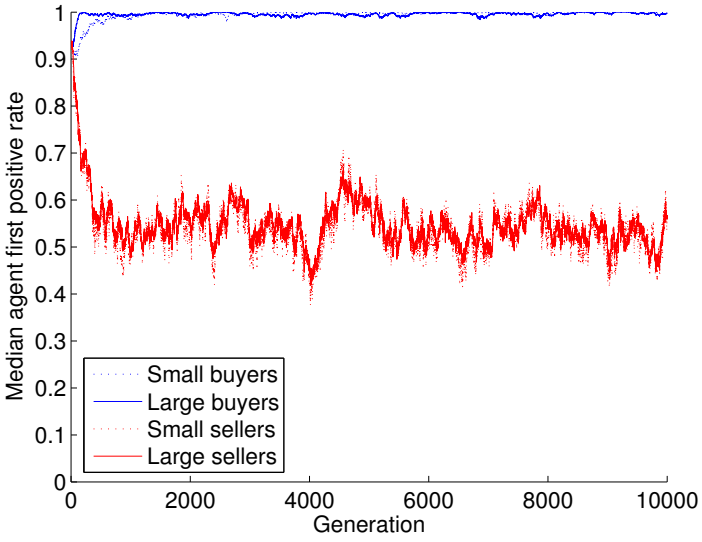


Figure 3.19: Evolution of the FP parameter in markets where users are forced to leave first negative feedbacks. Mean of 20 independent runs.

Stable Non-cooperative Markets

Of the 20 market simulations we ran, 20% evolve into a stable, non-cooperative state. One exemplar of this type of market is given in Figures 3.20–3.25.

Overall, the state of markets that become non-cooperative resembles that of the markets where retaliation is allowed. Participation in the reputation system is low, with agents unwilling to leave the first feedback, whether positive or negative. However, the state of the market appears slightly less dire than markets where reputation is allowed. During the course of a run, the "spikes" of increased cooperation and/or reputation system participation are both more frequent and longer in duration than similar periods seen in markets with reputation.

These four non-cooperative markets exhibit the effects of agent apathy overwhelming the positive benefits of using the reputation system. Removing retaliation removes one major disincentive to participation in the reputation system, but does nothing to actually encourage users to contribute. In most cases without retaliation, agents realize the benefits of cooperating and using the reputation system; nonetheless, merely removing retaliation does not appear to be sufficient to guarantee cooperation.

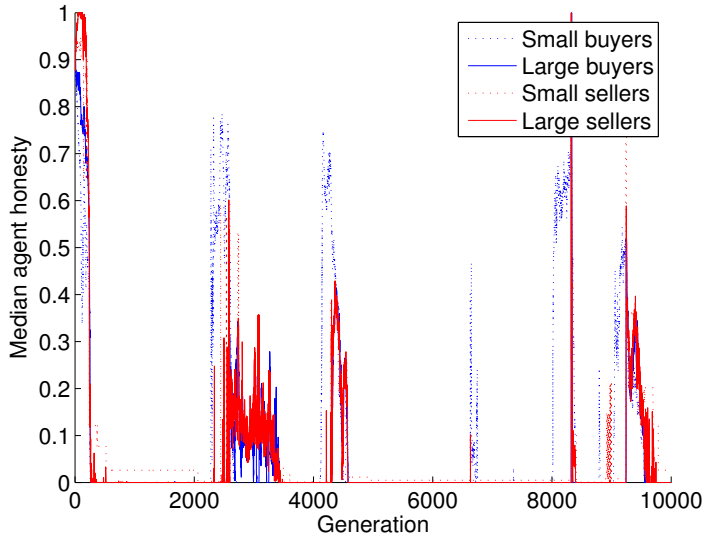


Figure 3.20: Evolution of the HON parameter in an example market without retaliation that remains in a stable, non-cooperative state.

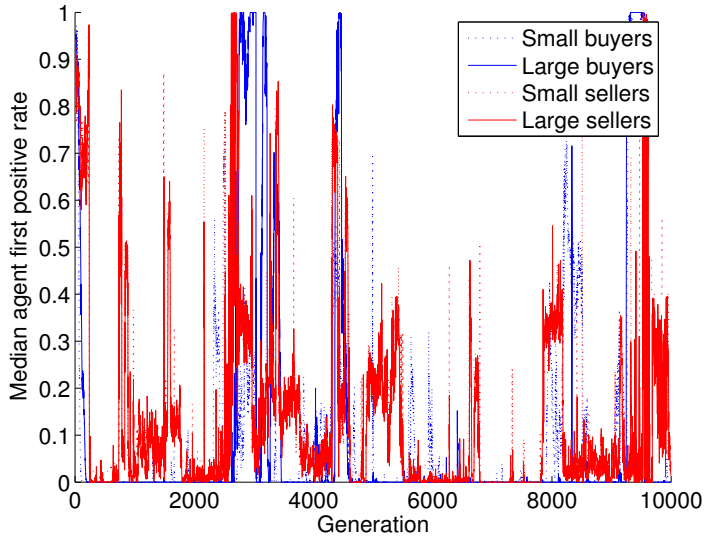


Figure 3.21: Evolution of the FP parameter in an example market without retaliation that remains in a stable, non-cooperative state.

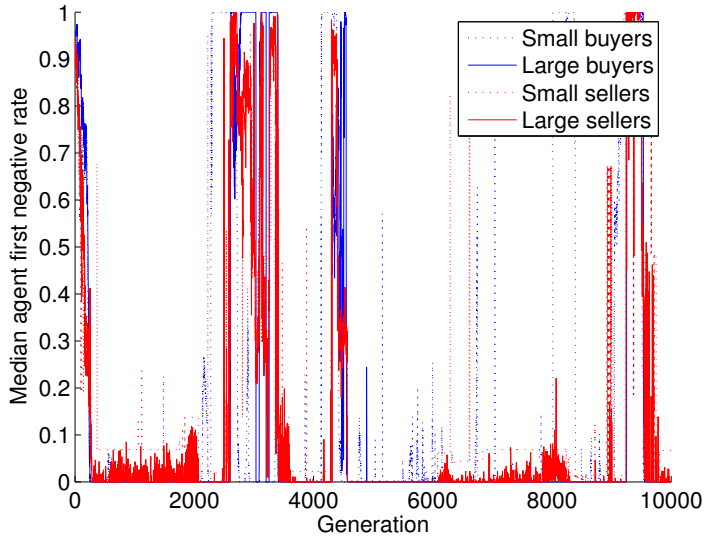


Figure 3.22: Evolution of the FN parameter in an example market without retaliation that remains in a stable, non-cooperative state.

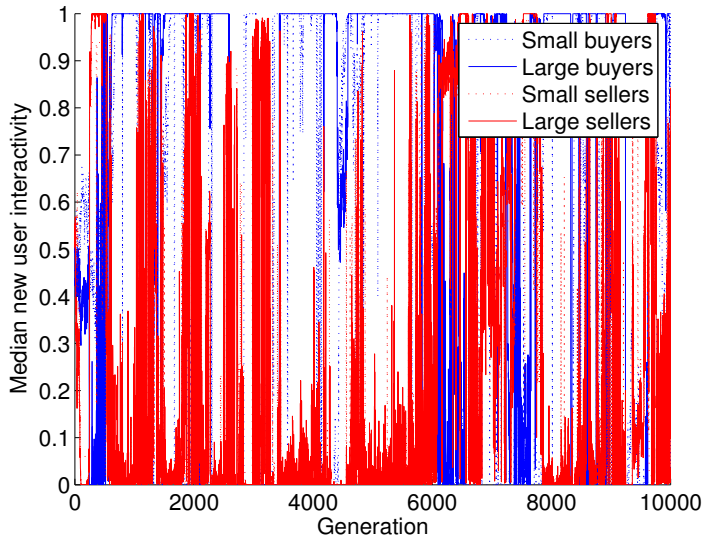


Figure 3.23: Evolution of the NUI parameter in an example market without retaliation that remains in a stable, non-cooperative state.

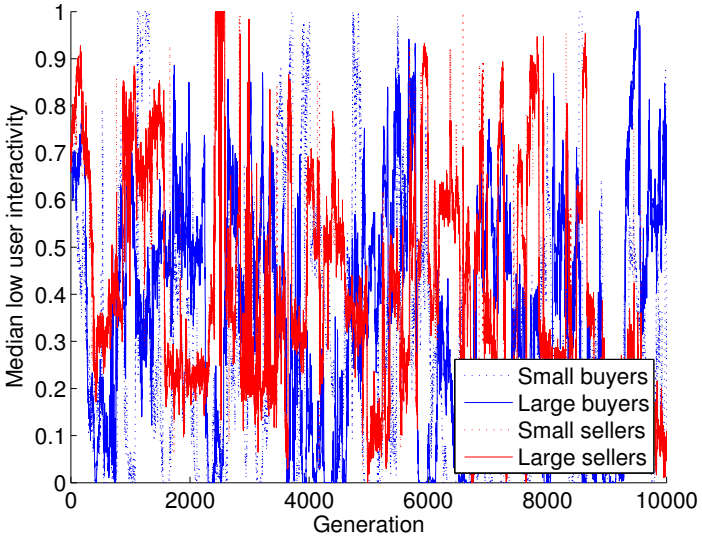


Figure 3.24: Evolution of the LUIT parameter in an example market without retaliation that remains in a stable, non-cooperative state.

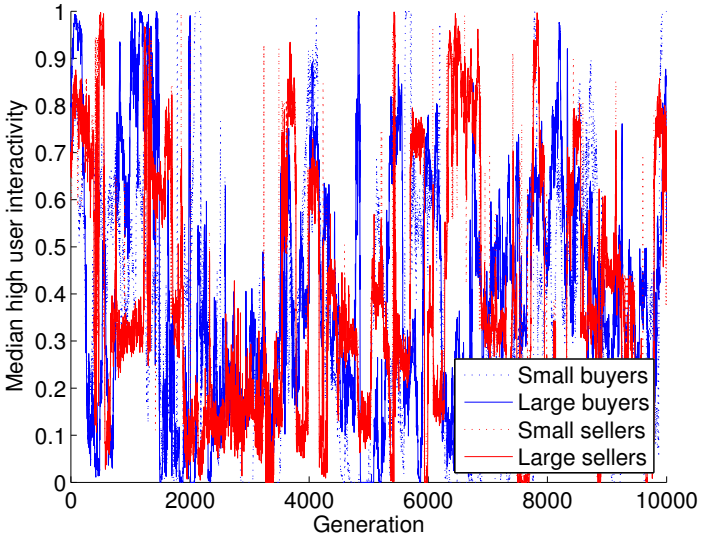


Figure 3.25: Evolution of the HUIT parameter in an example market without retaliation that remains in a stable, non-cooperative state.

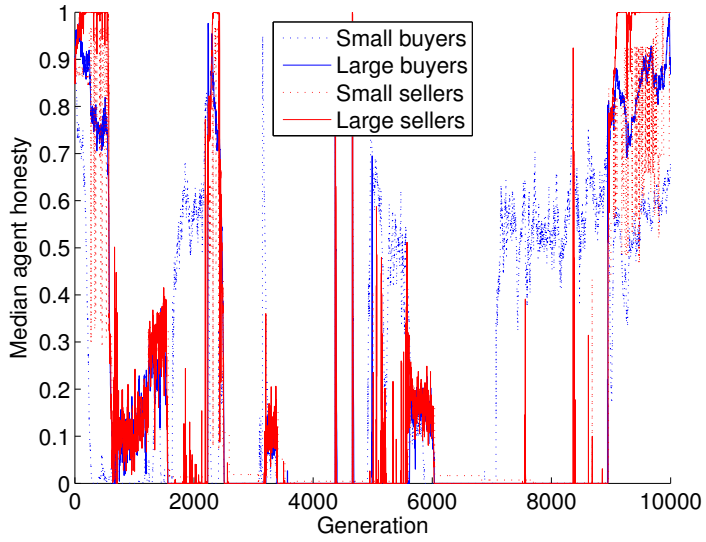


Figure 3.26: Evolution of HON parameter in Market A, one of the two markets that remain stable.

Unstable States

The remaining two markets exhibit perhaps the most interesting behavior: they do not remain in either stable state for the entire duration of the simulation. The first of these markets ("market A," Figures 3.26–3.31) initially appears to be in a cooperative state until around generation 500, at which point it rapidly degenerates into a non-cooperative state until generation 8950, where it just as suddenly flips back into cooperation. The second market ("market B," not shown) becomes non-cooperative at around the same point (generation 500), but remains non-cooperative for less than 500 generations before returning to cooperation.

While the small sample size and complexity of the interaction/transaction/ reputation games makes it difficult to state with confidence exactly what mechanisms cause a switch between cooperative and non-cooperative states, we can observe some common patterns in these two cases. In both cases, the small buyers very quickly become uncooperative, in contrast to stable cooperative markets like Figure 3.12, where small buyer cooperation remains above 0.75. In addition, small sellers have lower interaction thresholds than usual before the crash in these two markets. Finally, we observe a slow but steady decline in large buyer honesty leading up to the sudden collapse.

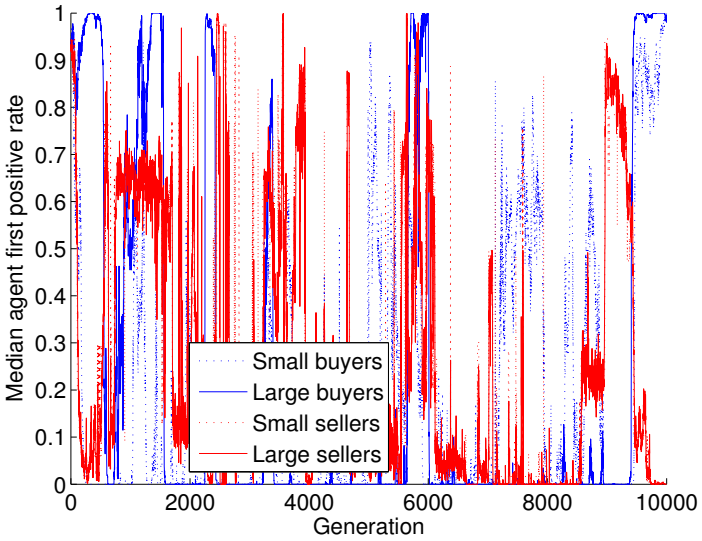


Figure 3.27: Evolution of FP parameter in Market A, one of the two markets that remain stable.

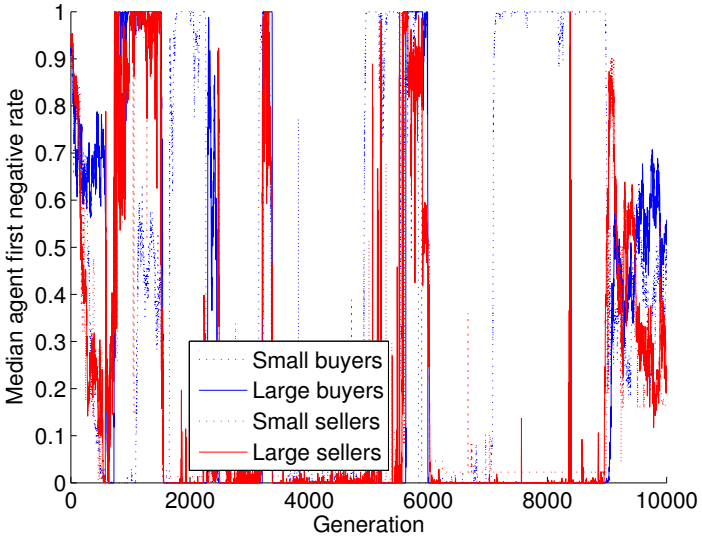


Figure 3.28: Evolution of FN parameter in Market A, one of the two markets that remain stable.

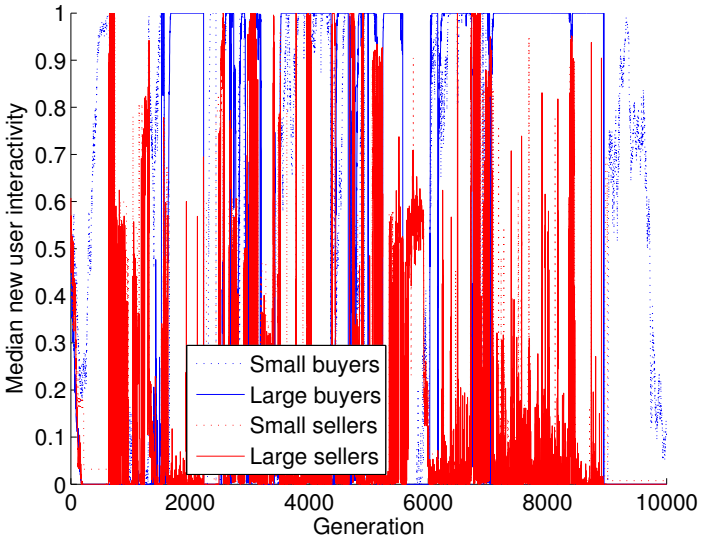


Figure 3.29: Evolution of NUI parameter in Market A, one of the two markets that remain stable.

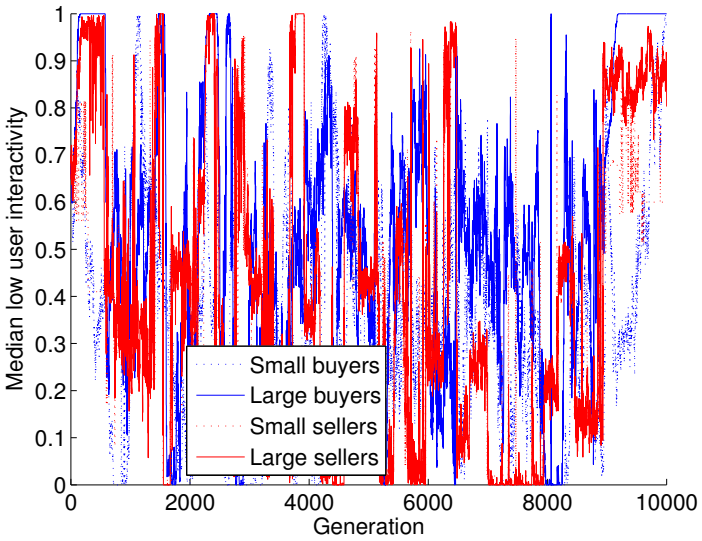


Figure 3.30: Evolution of LUIT parameter in Market A, one of the two markets that remain stable.

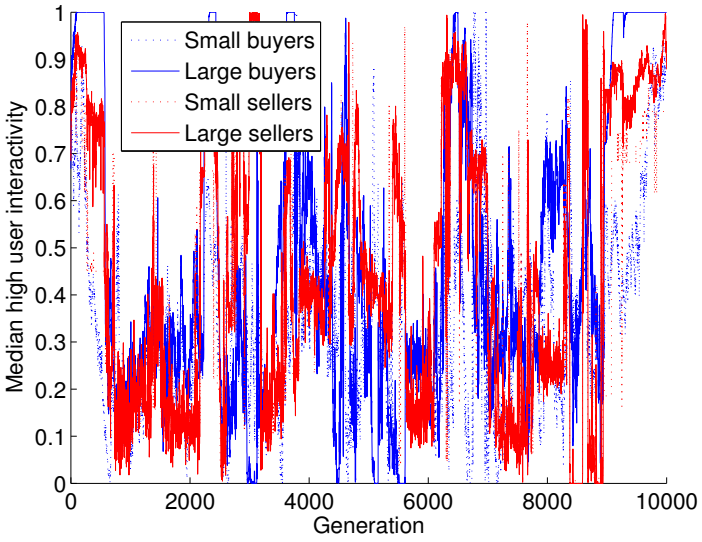


Figure 3.31: Evolution of HUIT parameter in Market A, one of the two markets that remain stable.

One plausible explanation is that the small buyers evolve a non-cooperative strategy quickly to take advantage of the relatively low interaction thresholds that exist because of the initial marketplace setup. The small sellers, at a disadvantage due to their low interaction rate, adopt lower interaction thresholds in attempt to balance the benefits of easier interaction against the risk of dealing with more dishonest sellers. Sensing an opportunity, the large buyers start cooperating less. Eventually, the buyer cooperation rate reaches a critical point beyond which it is no longer profitable for sellers to cooperate, and the entire market rapidly switches to non-cooperation.

The mechanism behind the return to cooperation is harder to explain. In both cases, it appears to be driven by increased *use* of the reputation system. Before the return to cooperation, both markets experience periods where first the seller, then the buyer median interaction thresholds increase. These periods of increased selectivity may apply some pressure on agents to behave honestly.

The sources of these interactivity threshold increases remain elusive. There is no clear motivator in market B, where the period of non-cooperation is so brief. In market A, the small buyers appear to again play a pivotal role. They start to cooperate nearly 2000 generations before any other agent class. Roughly 1000 generations after the small buyers start to cooperate, large seller LUIT starts

to increase. Then we see large buyer interaction thresholds begin to rise along with a fall in seller thresholds. Finally, the honesty rates of all agents jumps back to the cooperative state.

While it is difficult to completely characterize why these two markets fall into and out of a cooperative state, these results do provide evidence that neither state is completely stable. They also suggest that along with honest behavior and contribution of feedback, the use of reputation system data also play a part in keeping markets cooperative. After all, if users are not relying on the reputation system while playing the interaction game, there is little value to a good reputation.

It is encouraging that the only two markets to become non-cooperative after an initial period of cooperation did so during the initial period of evolution when parameters are still converging to a steady state. Likewise, that market A became cooperative again after over 8000 generations of non-cooperation suggests that if agents can overcome their apathy toward the reputation system, then cooperation is ultimately achievable. Clearly, further study is needed both to determine the mechanisms involved in the transition between cooperation and non-cooperation and to determine whether the remaining non-cooperative markets can make similar returns to cooperation.

3.4.3 Simultaneous Feedback

In our simulated market, it is easy to magically remove retaliation; real markets are not so accommodating. One straightforward approach to dealing with retaliation is to change the reputation game from a dynamic to a simultaneous game. In this model, the feedback process becomes a sort of secret ballot: both players make their feedback choice blind. Neither player's feedback is revealed to the other until both sides have either left feedback or committed to not leaving feedback.

There may still be ways to game the reputation system – for example, a player that defects may leave a "pre-emptive" negative to attempt to disguise his or her defection – but retaliation is impossible in such a system. Furthermore, blind feedback is a minor change to existing systems that would not require modifications to the feedback interface nor to the interpretation of existing reputations.

Our final experiment attempts to answer the question: does a reputation system that uses blind feedback (and thus cannot have retaliation) ensure the market evolves to a cooperative state? Once

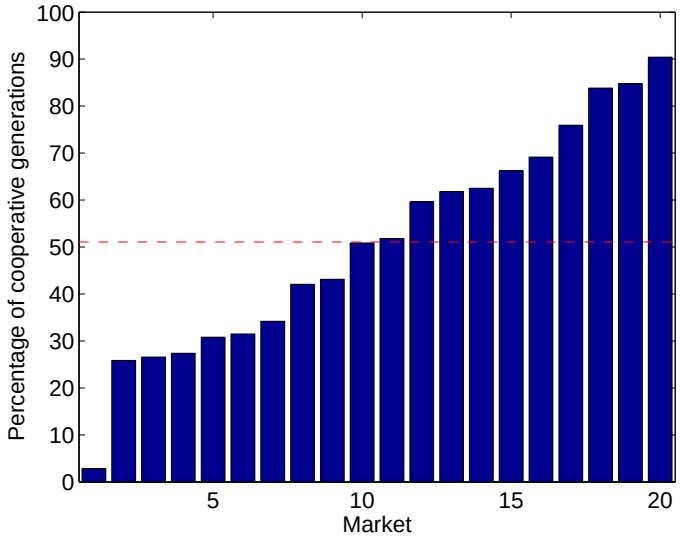


Figure 3.32: Percentage of time each market spends in a cooperative state during generations 501–10,000 with blind feedback. The dashed line indicates the mean cooperation rate across all 20 markets.

again, we simulate each of 20 markets out to 10,000 generations. In one of these markets, cooperation never evolves. The remaining 19 spend at least some of the time cooperating but all oscillate between a cooperative and uncooperative state. Figure 3.32 plots the fraction of time each of the 20 markets spent in a cooperative state. The mean market is cooperative 50.9% of the time while the best market cooperates 90.4% of the time. The worst market, aside from the completely uncooperative one, spends only 25.8% of its time cooperating.[2]

The evolution of agent honesty for four example markets (the best, the one closest to the mean, the worst that spends some time cooperating, and the completely uncooperative market) is shown in Figures 3.33–3.36. Once again, the exact mechanism for the collapse is difficult to determine, but agent apathy seems to be the main factor. In all these markets, participation in the reputation system is somewhat low: 0.4-0.5 on average. The rises and falls in agent honesty also occur with respective falls and rises in interaction threshold values, suggesting that in cooperative markets, agents become less selective during the interaction game, which allows defectors to infiltrate. When

[2] All of these percentages are calculated using generations 501–10,000. Because we start the simulations in a cooperative state, the initial period is cooperative in all markets. We disregard this period to avoid it biasing the results.

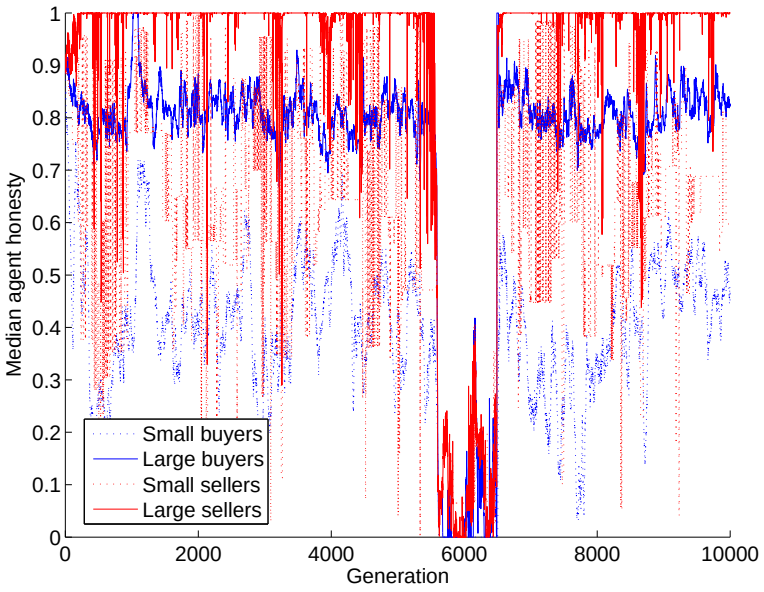


Figure 3.33: Evolution of agent honesty with blind feedback in an example market with high cooperation in 90.4% of generations.

agents in a non-cooperative state start to become more selective, cooperation once again becomes practical.

One lingering question is: why are the results from this experiment so different from the one with retaliation disabled directly? The answer appears to lie with our parameterization of player strategies. In the original dynamic reputation game, once one player decides to leave a first feedback, the other player always leaves a second feedback. In this experiment, the reputation game is simultaneous, so the two players independently decide whether to leave feedback. More formally, if the players' first feedback rates are both p, then the expected number of feedbacks per transaction is $4p - 2p^2$ in the original dynamic game and $2p$ with blind feedback. So for all possible values of p, we expect more feedback in the original dynamic formulation of the reputation game.

While eliminating this simulation bias (perhaps through a more sophisticated parameterization of player strategies) would be desirable, the results as they are lend further support to the conclusion that user apathy is a major stumbling block for this type of reputation system. The lower feedback rates in the simultaneous reputation game result in markets that are less likely to remain in a cooperative state.

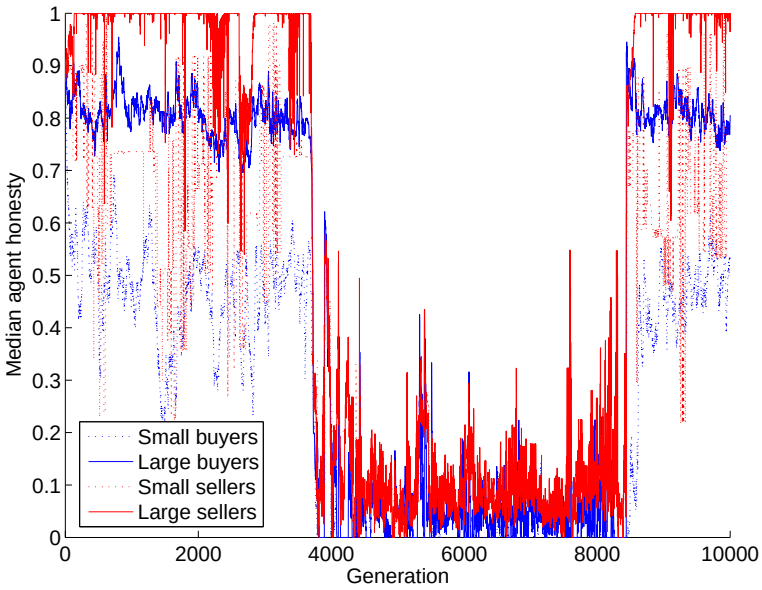


Figure 3.34: Evolution of agent honesty with blind feedback in an example market with high cooperation in 50.9% of generations.

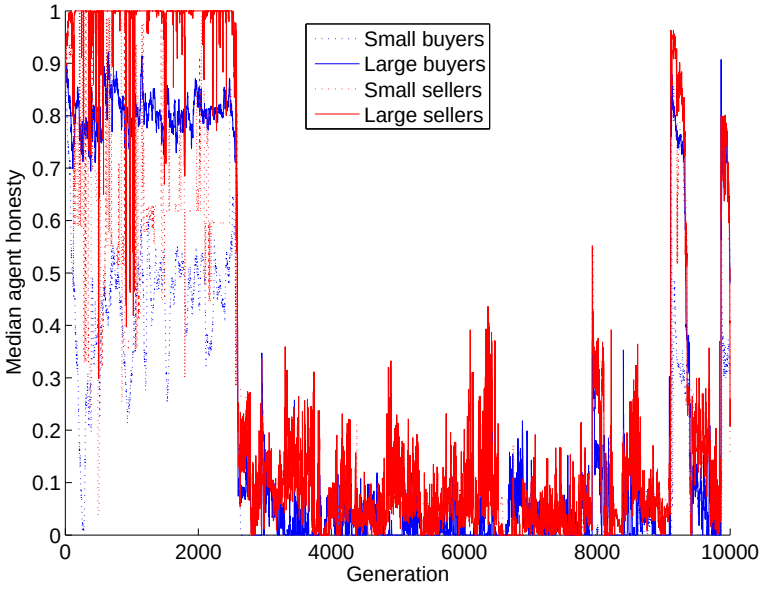


Figure 3.35: Evolution of agent honesty with blind feedback in an example market with high cooperation in 25.8% of generations.

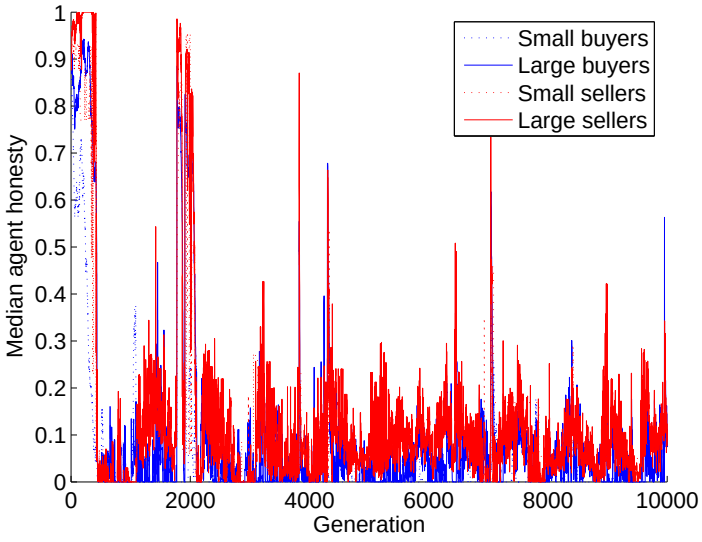


Figure 3.36: Evolution of agent honesty with blind feedback in an example market with high cooperation in 2.8% of generations.

Along with the previous experiments, these results demonstrate that a retaliation-free reputation system is a necessary, but not sufficient, condition for a cooperative marketplace. User apathy, both in leaving and using feedback, can still cause the collapse of cooperation even when all feedback given is honest.

3.5 Conclusion

In this chapter, we propose a simple model for peer-to-peer markets that nonetheless exhibits many of the complexities of real systems. At the heart of this model are two widely studied games: the Prisoners' Dilemma and the tragedy of the commons. This formulation allows us to model many of the observed subtleties of the trading and feedback process and allows us to draw the following conclusions:

Reputation systems can maintain cooperation. In an ideal setting, such as when users are forced to leave accurate feedback at a high rate, a simple reputation system like PPF is sufficient for maintaining a cooperative market. When we force agents to leave first negative feedback

(Figures 3.18 and 3.19), first positive rates are also high and all of the simulated markets remain cooperative.

Selfish behavior drastically reduces reputation system performance. When we let agents find optimal strategies for participating in the reputation system, enough chose to retaliate that the reputation system is no longer able to maintain cooperation in any of the markets. Similar patterns of retaliation and the consequent decline in negative feedback are commonly observed in real markets, suggesting that existing reputation systems are not realizing their full potential, even though other factors appear to be preventing complete non-cooperation.

Encouraging user participation is essential. Even with retaliation prohibited, many users choose not to participate in the feedback process. While not unexpected — in most reputation systems there is neither a reward for participating nor a punishment for abstaining — user apathy can also cause a collapse of cooperation. The results of Sections 3.4.2 and 3.4.3 suggest that it is not enough to eliminate retaliation and other disincentives to participation in the reputation system. To fully ensure cooperation, we need a mechanism that rewards frequent, honest contributions of feedback from all users.

These results are generally intuitive, but are now backed-up by an underlying theory and simulation. Moreover, particulars of the details are perhaps unexpected, such as the high rate of agent apathy even when retaliation is disabled and there is no penalty for leaving feedback first.

Real markets are undoubtedly more complex than this model. Nevertheless, we believe it is useful to generalize the agent behaviors observed in our model and use it to help design reputation systems for real markets. In particular, our results suggest that successful reputation systems will need to both ensure that there is no punishment for leaving negative feedback and that honest participation in the reputation system, a certain social benefit, is rewarded.

Chapter 4

Mitigating Retaliation with EM-Trust

4.1 Introduction

As we showed in the preceding chapter, retaliation for negative feedback can have a powerfully adverse effect on the functioning of a reputation system. In addition to reducing the accuracy of the recipient's reputation, retaliatory negative feedback also produces a chilling effect on participation in the feedback process. Some users — typically large sellers with a large interest in protecting their reputations — wield the threat of retaliation, which leads to a systematic underreporting of failed transactions. (95) Some researchers believe that the market owners ignore this phenomenon because a marketplace with abundant positive feedback appears safer and more inviting to new customers. (16)

In this chapter, we discuss strategies for mitigating these effects. We first propose EM-trust, an algorithm that uses a latent variable model of the feedback process to estimate reputations using Expectation-Maximization. EM-trust essentially assumes away the problem of retaliation by distributing fault for a failed transaction in a statistically fair manner.

Using our marketplace simulator, we show that EM-trust estimates users' true reliability with accuracy approximately equal to that of Percent Positive Feedback (PPF) in the absence of retaliatory negative feedbacks, but that with high rates of retaliation, EM-Trust returns significantly more accurate reputations. These more accurate reputations have a direct effect on the performance of

the market: marketplaces using EM-Trust are just as safe as ones using eBay's PPF system — the rate of failed transactions is essentially the same — but the EM-Trust market deactivates fewer good users unfairly and has higher liquidity.

We also address the problem of data sparseness prevalent in peer-to-peer reputation systems by proposing Bayesian variants of both EM-trust and eBay's PPF. Since the prior distribution of user reliability needed for a Bayesian estimator is not immediately available, we demonstrate how a workable prior distribution can be estimated on-line from feedback data. Bayesian estimation benefits both EM-trust and PPF by improving reputation accuracy when tested with a realistically sparse dataset. The combination of the latent-variable EM-Trust algorithm augmented with a Bayesian prior effectively manages both missing and inaccurate feedback data and produces the most accurate estimates of users' true reliability.

4.2 Algorithm Design

We begin by defining a user i's reliability, λ_i, as the probability that the user will perform acceptably in a transaction. As suggested by (8), we currently do not try to assess motivation — poor performance caused by deception or malice is treated indistinguishably from mere incompetence. Since the end result is the same, we do not think it is necessary to discriminate between sources of unacceptable performance. Acceptable performance for a seller means selling only accurately described, functional products and sending the goods in a timely fashion through a reliable shipper. For a buyer, it involves remitting payment in an approved form on time.

The job of the reputation system, then, is to accurately estimate λ_i from users' (possibly unreliable) feedback. When user i interacts with user j in a transaction, we observe two feedback variables, F_{ij} and F_{ji}, indicating the feedback left by user i for user j and vice versa. These variables are can take values from the set $\{-1, 1, 0\}$ indicating negative, positive, and no feedback respectively. We do not currently model neutral feedback, since it is both infrequently given and is considered by most to be merely a weak negative.

Also associated with each transaction are two latent Bernoulli random variables, T_{ij} and T_{ji}. T_{ij} indicates whether user i performed acceptably in the transaction with user j, and T_{ji} represents

user j's performance in the same transaction. We assume independence between transactions, so the distribution of these Bernoulli random variables is characterized by the users' reliability parameters, λ_i and λ_j.

In general, the feedback variables can depend on the individual performance variables as well as on each other. These dependencies are complex and difficult to quantify, so we do not attempt to model them explicitly. However, we do make some assumptions about the way rational users leave feedback. Our first assumption is that all transactions are legitimate transactions between real users. Reputation fraud, such as ballot stuffing and bad mouthing, perpetrated by fake transactions and sybil users is an interesting problem in its own right, which we discuss in Chapter 5.

Second, we assume that positive feedback always indicates that the recipient behaved acceptably in the transaction. While users may leave positive feedback despite small faults (e.g. slow shipping) in the hope of encouraging a reciprocal positive, it seems unlikely that a user would leave positive feedback for a grossly under-performing partner just to get a positive in return. In any case, we cannot discern a false positive from a true positive feedback, so we consider them all to be legitimate.

Finally, we assume that a negative feedback by itself does not indicate poor performance, unless the recipient of the negative feedback has left a positive for his or her partner. We know that the process of retaliation creates false negatives, and we also hypothesize that nefarious users may leave pre-emptive false negatives to try to disguise bad behavior.

This model is illustrated graphically in Figure 4.1. The top layer consists of the user reliability parameters λ_i and λ_j. In the standard EM-Trust formulation, these are treated as parameters while in the Bayesian version they are latent random variables. The middle layer is the latent transaction success indicator variables. The bottom layer is the observed multinomial feedback variables. Arrows indicate conditional dependencies between the terms in this model. Retaliatory negative feedback arises in this model as a consequence of the second feedback's (F_{ji}) dependence on the first feedback (F_{ij}).

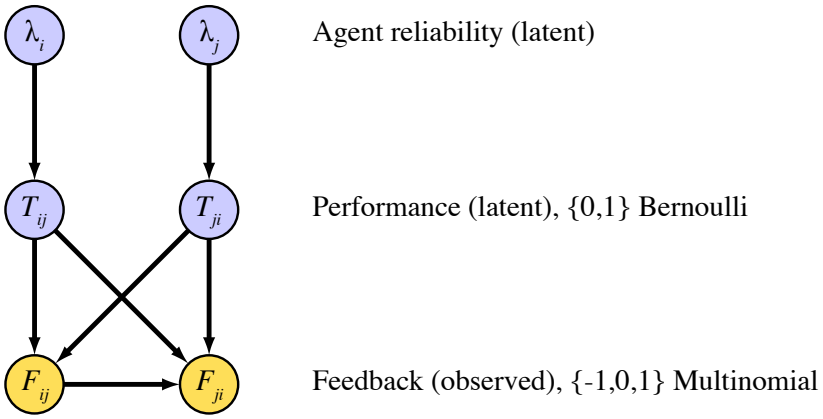


Figure 4.1: Graphical model illustrating the EM-Trust model for a single transaction.

4.2.1 Percent Positive Feedback: A Baseline Reputation System

In order to evaluate the effectiveness of our new reputation algorithms, we must first define a baseline system against which we can measure any differences. For this series of experiments, we use Percent Positive Feedback (PPF), one component of the eBay Feedback Forum, as our baseline. For each user in the market, the Feedback Forum provides the PPF score as well as a Feedback Score, the number of feedbacks received minus the number of negatives, and a log of every feedback received along with a short comment from the giver.

For a user i who has received feedback from a set of other users A, PPF's estimate of the user's underlying reliability is trivially easy to compute:

$$\hat{\lambda}_i^{(\text{PPF})} = \frac{1}{|A|} \sum_{j \in A} \mathbb{1}_{F_{ji}=1}$$

PPF is essentially a simple mean of feedback received. The other aggregated reputation metric, the Feedback Score, can be used to determine the sample size, $|A|$, allowing users to estimate their confidence in the PPF's estimate.

PPF treats all feedback as though it were honest. It makes no attempt to determine whether or not a negative is justified or simply due to retaliation.

4.2.2 The EM-trust Algorithm

EM-Trust improves on PPF by using the full latent model outlined above. If we could observe T_{ij} and T_{ji}, estimating λ_i and λ_j would be trivial. However, these variables are latent, so we proceed by using an Expectation-Maximization (EM)algorithm to iteratively refine our estimates of these parameters. We start with initial estimates $\hat{\lambda}_i^{(0)}$ for each user's reliability parameter. These starting values do not have a great effect in our application, so we start with $\hat{\lambda}_i^{(0)} = 0.5$ for all i. The EM algorithm then alternates between two phases. During the expectation step (E-step), it uses the current estimates of the parameters to calculate conditional expectations of the latent variables. Then, in the maximization step (M-step), it calculates estimates of the parameters using the conditional expectations computed during the E-step. The process then repeats until the parameter estimates converge.

Expectation Step

For each transaction involving user i, we calculate the conditional expectation that i performed acceptably given the observed feedback:[1]

$$\begin{aligned}
\mathbb{E}[T_{ij}|F_{ij}=1,F_{ji}=1] &= 1 \\
\mathbb{E}[T_{ij}|F_{ij}=0,F_{ji}=1] &= 1 \\
\mathbb{E}[T_{ij}|F_{ij}=-1,F_{ji}=1] &= 1 \\
\mathbb{E}[T_{ij}|F_{ij}=1,F_{ji}=-1] &= 0 \\
\mathbb{E}[T_{ij}|F_{ij}=0,F_{ji}=-1] &= \mathbb{E}[T_{ij}|T_{ij}T_{ji}=0] \\
\mathbb{E}[T_{ij}|F_{ij}=-1,F_{ji}=0] &= \mathbb{E}[T_{ij}|T_{ij}T_{ji}=0] \\
\mathbb{E}[T_{ij}|F_{ij}=-1,F_{ji}=-1] &= \mathbb{E}[T_{ij}|T_{ij}T_{ji}=0]
\end{aligned}$$

We do not compute the expectation for the two unlisted cases ($F_{ij}=1, F_{ji}=0$ and $F_{ij}=$

[1] We observe a slight increase in accuracy if we let $\mathbb{E}[T_{ij}|F_{ij}=0,F_{ji}=-1]=0$ and treat $\mathbb{E}[T_{ij}|F_{ij}=-1,F_{ji}=0]$ as missing data. However, doing so creates an incentive for retaliation beyond the social incentives that already exist. Thus, rational users will always retaliate, so performance with this modification should converge to that of the version presented in the body of this report. Since our simulated users do not adapt to the system being tested, we feel any performance gains from this modification are due to simulation biases rather than a true improvement in accuracy. However, when deploying EM-trust in a real market, it may be worthwhile to reconsider this modification.

$F_{ji} = 0$) and instead treat these transactions as missing data. The calculations for $\mathbb{E}[T_{ji}|F_{ij}, F_{ji}]$ are the same with the subscripts reversed.

Because we assume positive feedback is always accurate, the user who received a positive feedback is known to be certainly honest, so $\mathbb{E}[T_{ij}|F_{ji} = 1] = 1$ in this case. Similarly, if user i left positive feedback for user j but received negative feedback, the negative feedback cannot be retaliation, so it must indicate the i behaved unacceptably.

When at least one negative and no positive feedback is given, all we know is that someone behaved unacceptably. We cannot rely on the feedback being accurate in these cases, so we use the more fundamental expectation $\mathbb{E}[T_{ij}|T_{ij}T_{ji} = 0]$. To compute this expectation, we look at the joint distribution, $\mathbb{P}\{T_{ij}T_{ji}, T_{ij}, T_{ji}\}$, marginalize over T_{ji}, and condition on $T_{ij}T_{ji} = 0$:

$$\begin{aligned}
\mathbb{E}[T_{ij}|T_{ij}T_{ji} = 0] &= \mathbb{P}\{T_{ij} = 1|T_{ij}T_{ji} = 0\} \\
&= \sum_{t=0,1} \frac{\mathbb{P}\{T_{ij} = 1, T_{ji} = t, T_{ij}T_{ji} = 0\}}{\mathbb{P}\{T_{ij}T_{ji} = 0\}} \\
&= \frac{\mathbb{P}\{T_{ij} = 1, T_{ji} = 0\}}{\mathbb{P}\{T_{ij}T_{ji} = 0\}} \\
&= \frac{\hat{\lambda}_i^{(t)}(1 - \hat{\lambda}_j^{(t)})}{1 - \hat{\lambda}_i^{(t)}\hat{\lambda}_j^{(t)}}
\end{aligned}$$

This estimation process is the key to the EM-trust algorithm. We assume that negative feedback is mostly unreliable and so we penalize both parties in such a transaction. The amount of "blame" given to each is based on the current reputations of the two parties in a transaction, with more blame given to the lower reputation user. The aim of this technique is to render retaliatory feedback irrelevant. If a user has received a negative feedback, it does not matter whether he or she leaves a retaliatory negative or not: the reputations of both parties will be computed in the same fashion regardless of whether the user retaliates.

The only way in which the recipient of a negative feedback can change the outcome of the reputation process is by leaving a positive for its partner, which will have the effect of shifting *all* blame for the transaction failure onto himself, lowering his reputation even further. On its face, it is not desirable for the reputation system to discourage users from leaving honest feedback. However, users rarely leave positive feedback for others that gave them negative feedback, even when such behavior will not result in a lower reputation: Resnick and Zeckhauser (95) report that none of the

buyers and only 13% of sellers who received negative feedback respond with a positive feedback. Therefore we feel that making slightly more optimal the pre-existing strategy of not praising those who criticize you is a worthwhile trade-off for mitigating the effect of the far more damaging tactic of retaliation.

Maximization Step

For the maximization step, we use the above conditional expected values as the sufficient statistics needed to update estimates of the parameters λ_i. Let $\langle T_{ij} \rangle^{(t)} = \mathbb{E}[T_{ij}|F_{ij}, F_{ji}]$ be the conditional expectation of user i's behavior in a transaction with user j computed in the expectation step of iteration t. Let A be the set of users with which user i has interacted and for whom we have computed $\langle T_{ij} \rangle^{(t)}$. We estimate the updated value of user i's reliability parameter as

$$\hat{\lambda}_i^{(t+1)} = \frac{1}{|A|} \sum_{j \in A} \langle T_{ij} \rangle^{(t)} \tag{4.1}$$

We then use these updated parameter estimates in the next iteration's E-step. We continue this process until the estimates converge, and take values of the final iteration, T, as our parameter estimates:

$$\hat{\lambda}_i^{(\text{EM-Trust})} = \hat{\lambda}_i^{(T)}$$

If a pair of users have had multiple transactions together, EM-trust behaves like PPF and only counts the most recent transaction, making it more difficult for malicious users to create bogus reputations by leaving multiple positive feedbacks for sales of non-existent items. If it should become desirable to include all transactions between two users, the modifications would be trivial.

As an iterative algorithm, EM-trust is naturally more expensive to run than PPF. Each iteration, though, is linear in the number of transactions, and convergence is fast in practice. One interesting consequence of this iterative process is that future negative feedbacks can affect the amount of blame one receives for past failed transactions. The value of $\mathbb{E}[T_{ij}|T_{ij}T_{ji} = 0]$ depends on both parties' reliability estimates, which depend of *all* the feedback they have received at the time EM-Trust is run, including feedback for transactions that occurred after T_{ij}. Therefore, if a user's reputation falls due to new negative feedback, subsequent reputation evaluations will increase the amount of blame

assigned to the user for past failed transactions, further decreasing the user's reputation. While a second order effect, we believe this process to be beneficial because it provides yet another incentive to maintain a high reputation.

4.2.3 Bayesian Estimation

To help manage data sparseness, we also derive Bayesian versions of PPF and EM-Trust that use a prior distribution of user behavior as well as the observed data to estimate the user's actual behavior distribution. For a prior we use a mixture of Beta distributions:

$$\gamma \text{Beta}(\alpha_1, \beta_1) + (1 - \gamma)\text{Beta}(\alpha_2, \beta_2) \tag{4.2}$$

where the α_1 and β_1 parameters describe the probability distribution of acceptable performance among users that are mostly honest and competent (i.e., the reliable users) and α_2 and β_2 describe the distribution of acceptable performance among mostly dishonest or incompetent (i.e. unreliable) users. The γ parameter describes the proportion of reliable users in the market. As with all Bayesian methods, estimating these parameter values is one of the main challenges for a successful implementation.

For EM-trust, the estimation step remains the same, but the maximization step replaces the simple maximum likelihood estimate with the mean of the posterior distribution:

$$\hat{\lambda}_i^{(t+1)} = \gamma' \frac{\alpha_1'}{\alpha_1' + \beta_1'} + (1 - \gamma') \frac{\alpha_2'}{\alpha_2' + \beta_2'}$$

where

$$\begin{aligned}
\gamma' &= \frac{\gamma B(\alpha_1', \beta_1') B(\alpha_2, \beta_2)}{\gamma B(\alpha_1', \beta_1') B(\alpha_2, \beta_2) + (1 - \gamma) B(\alpha_2', \beta_2') B(\alpha_1, \beta_1)} \\
\alpha_1' &= \alpha_1 + \sum_{j \in A} \langle T_{ij} \rangle^{(t)} \\
\beta_1' &= \beta_1 + |A| - \sum_{j \in A} \langle T_{ij} \rangle^{(t)} \\
\alpha_2' &= \alpha_2 + \sum_{j \in A} \langle T_{ij} \rangle^{(t)} \\
\beta_2' &= \beta_2 + |A| - \sum_{j \in A} \langle T_{ij} \rangle^{(t)}
\end{aligned}$$

and $B(\alpha, \beta)$ is the Beta function, $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$. This posterior distribution of the λ_i is also a mixture of Betas with parameters α_1', β_1', α_2', β_2', and γ'.

We use a similar approach to create a Bayesian version of PPF:

$$\hat{\lambda}_i^{\text{(Bayesian-PPF)}} = \gamma' \frac{\alpha_1'}{\alpha_1' + \beta_1'} + (1 - \gamma') \frac{\alpha_2'}{\alpha_2' + \beta_2'}$$

where γ' is defined as above and

$$\begin{aligned}
\alpha_1' &= \alpha_1 + \sum_{j \in A} \mathbb{1}_{F_{ji}=1} \\
\beta_1' &= \beta_1 + |A| - \sum_{j \in A} \mathbb{1}_{F_{ji}=1} \\
\alpha_2' &= \alpha_2 + \sum_{j \in A} \mathbb{1}_{F_{ji}=1} \\
\beta_2' &= \beta_2 + |A| - \sum_{j \in A} \mathbb{1}_{F_{ji}=1}
\end{aligned}$$

4.2.4 Estimating the Bayesian Prior

The Bayesian versions of EM-trust and PPF realize all of their advantages by incorporating knowledge of the prior distribution of user reliability during the estimation process. Without information about this prior distribution, these techniques are useless: using a non-informative (uniform) prior results in lower performance than the non-Bayesian versions of the algorithms.

We use an empirical Bayes method to estimate the prior using nothing but the observed feedback by exploiting the fact that non-Bayesian EM-trust and PPF both calculate acceptably accurate estimates of user reliability. We assume that the population reliability distribution has a binary mixture-of-Betas density, then estimate the parameters for this model from the non-Bayesian reputation values. While there are several methods for mixture-of-Betas density estimation, all of our results below use an EM approach, because compared to a direct gradient ascent algorithm, EM both runs faster and yields more accurate parameter estimates.

This estimated population reliability distribution becomes the prior in the Bayesian versions of EM-trust and PPF. During simulation, we periodically (every 10,000–25,000 transactions) re-estimate the prior by running the non-Bayesian reputation algorithm to generate reliability values and once again fit a density to the results.

As our results below show, the use of a prior has a large positive impact on the performance of both PPF and EM-Trust. In terms of both Kullback-Liebler divergence and empirical performance in the simulated market, the distributions estimated with this technique lie roughly halfway between the true distribution of user honesties and a non-informative uniform prior. Improved techniques for performing this estimation are one clear route to increased accuracy of all the algorithms discussed here.

4.3 Testing Methodology

Ideally, the best way to test these algorithms would be to conduct a controlled study comparing them to existing reputation systems in a actual functioning market. Unfortunately, such an experiment is not feasible.

We also consider testing using historical market data, but encounter several problems with this approach as well. Such information is typically proprietary and we are unable to procure a data set large enough to conduct our experiments. Furthermore, it is not clear that such testing would be entirely sufficient. Our goal is to craft a reputation system that affects user behavior — reducing the incentive retaliate and increasing willingness to leave legitimate negative feedback — so using only historical data would not measure the impact of the change in the reputation algorithm. Admittedly, it may provide a better snapshot of user interactivity patterns than our simulations, so we look forward to the possibility of validating our algorithms on historical transactions, should a source of data become available to us.

Because testing reputation systems in a real market is not feasible, we use a simulator to evaluate our algorithms. While simulated results are at best an approximation of what can be expected in a real marketplace, we believe it is a better demonstration of a reputation system's characteristics than the simple test cases or theoretical bounds that are prevalent in the literature. Simulation also permits us to test algorithms in multiple markets with varied characteristics in order to examine robustness across a range of plausible scenarios. Our simulator, used in most of the experiments in this thesis, is described in detail in Appendix A.

In order to demonstrate the effect of different marketplace characteristics on the reputation sys-

Parameter	Value
Proportion of "good" users (γ)	0.99
Good user α	198
Good user β	2
Bad user α	2
Bad user β	198
Initial number of buyers	8000
Initial number of sellers	1000
New user creation rate (% growth per day)	1.15
Respawn probability	0.60
Seller mean buy rate (items/day)	0.08
Seller mean sell rate (items/day)	0.8
Buyer mean buy rate (items/day)	0.2
Buyer mean sell rate (items/day)	0.008
Transactions per epoch	1000
Number of epochs	50
Good user probability of first feedback	0.40
Good user probability of second feedback	0.70
Bad user probability of first feedback	0.02
Bad user probability of second feedback	0.6

Table 4.1: Parameters for Market A, designed to have user types, interaction rates, and feedback patterns similar to those in eBay as observed by (95).

tems we test, we use two different simulated marketplaces. The parameters of our first marketplace — henceforth called *Marketplace A* — are chosen such that certain statistics, such as the rates of different types of feedback and the distribution of number of transactions per user, have values after 500,000 transactions that are close to the values reported by (95) in their study of eBay. The market begins with 8000 buyers and 1000 sellers. Of these users, 1% are "bad," with a mean reliability of 0.01. The remaining 99% are "good" and have mean reliability 0.99. The overall mean reliability in this market is thus 0.9802. New users are added at the rate of 1.15% per simulated day and have the same reliability distribution as the original set of users. Users that leave due to a low reputation return to the market 60% of the time, and their new identity keeps the reliability and other characteristics of their old identity. The specific parameters used in Market A are given in Table 4.1. While the reliability values in Marketplace A may seem remarkably high, they do agree with observations

Parameter	Value
Proportion of "good" users (γ)	0.98
Good user α	18
Good user β	2
Bad user α	2
Bad user β	18
Initial number of buyers	4000
Initial number of sellers	1350
New user creation rate (% growth per day)	0.25
Respawn probability	0.60
Seller mean buy rate (items/day)	0.08
Seller mean sell rate (items/day)	0.8
Buyer mean buy rate (items/day)	0.2
Buyer mean sell rate (items/day)	0.008
Transactions per epoch	1000
Number of epochs	50
Good user probability of first feedback	0.30
Good user probability of second feedback	0.60
Bad user probability of first feedback	0.1
Bad user probability of second feedback	0.5

Table 4.2: Parameters for Market B, a smaller market with a higher variance user reliability distribution.

that the vast majority of users at eBay behave correctly essentially all the time. However, some researchers believe that the extremely high reliability at eBay is due to systematic underreporting of minor faults in the Feedback Forum. To show how the reputation systems perform in a market with less than perfect users, we created *Marketplace B*. Marketplace B is slightly smaller (1350 sellers and 4000 buyers) and does not grow as fast (0.25% per day). However, the main difference is the distribution of user reliability. In Marketplace B, 98% of the users are "good," but the mean of good user reliability is only 0.9. Likewise, the bad users' mean reliability is 0.1, so the overall user mean is 0.884. This market models users that have higher variance around their mean behavior. By showing results in both of these marketplaces, we obtain some indication of how sensitive various reputation systems might be to the actual distribution of user reliability. The simulation parameters for Market B are shown in Table 4.2.

4.4 Experimental Results

In order to test EM-trust, we run a series of simulations and use several metrics to compare PPF, EM-trust, and their Bayesian variants. Our first test measures how accurately the four reputation systems can estimate the true underlying performance parameters for the users in a marketplace. In the second test, we evaluate how well the algorithms detect and eliminate low performance users. The final test measures the influence they have on the liquidity of the simulated market.

The Bayesian results presented in this section use priors estimated using the methods discussed in Sect. 4.2.4. We tested two density estimation methods for determining the prior: direct gradient ascent on the mixture model log likelihood and an EM mixture of Betas density estimator. Since both of these methods find only local maximum likelihoods, we run each estimation several times at random starting points. EM density estimation consistently outperforms direct gradient ascent and has the additional benefit of running faster. We found that around five random restarts achieves the optimal balance between estimate quality and run time.

The resulting distributions have a mean Kullback-Liebler divergence from the true prior of 2.6 in Market A and 1.1 in Market B, while the non-informative prior's divergence from the true prior in these markets is 5.3 and 1.9, respectively. As we show below, even this simple prior estimation method offers significant advantages over not using prior information. However, that there are still differences between the true priors and these estimates suggest that there are opportunities to further improve reputation system performance by simply refining prior estimation techniques. The performance when using the true prior gives an upper bound on these potential gains of about 25-35% in Market A and 35-45% in Market B.

4.4.1 Predicting Reliability

Our first experiment measures the reputation systems' abilities to learn the actual reliability of a set of users. For each reputation system, we run simulations at three different levels of retaliation. The 0% level simulates a marketplace where all feedback is completely accurate. At the 50% level, good users leave retaliatory negative feedback for about half of the negative feedbacks they receive, while bad users always retaliate. At the 100% level, all users always retaliate for negative feedback.

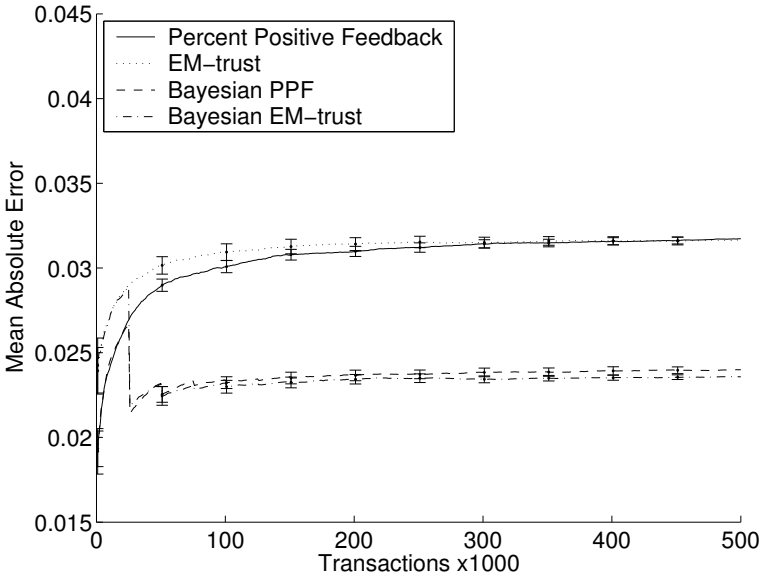


Figure 4.2: Reputation system mean absolute error in Market A with no retaliation.

From the data in (95) and assuming that simultaneous failure of both the buyer and seller is rare, the 50% retaliation level represents our best estimate of the rate of retaliation at eBay.

After each epoch of 1000 transactions, we compute the mean absolute error between the reputations returned by the reputation systems and the known ground truth reliability of the users in the system. The results of this test are shown in Figures 4.2–4.7.

With no retaliation, PPF and EM-trust perform roughly equally. In Market B, EM-trust is slightly more accurate, while in Market A, PPF is better, at least during the early stages of simulation. However, the advantages of EM-trust become apparent as the level of retaliation is increased. While both algorithms become less accurate with more retaliation, PPF's performance decreases more than EM-trust. This experiment demonstrates that both EM-trust variants accomplish the goal we set for them: they perform about as well as PPF in the zero retaliation case, and are much less influenced by inaccurate feedback data introduced by retaliatory negatives.

These results also demonstrate the power of incorporating prior information in the estimation process. The worse of the two Bayesian algorithms outperform both of the non-Bayesian ones. The

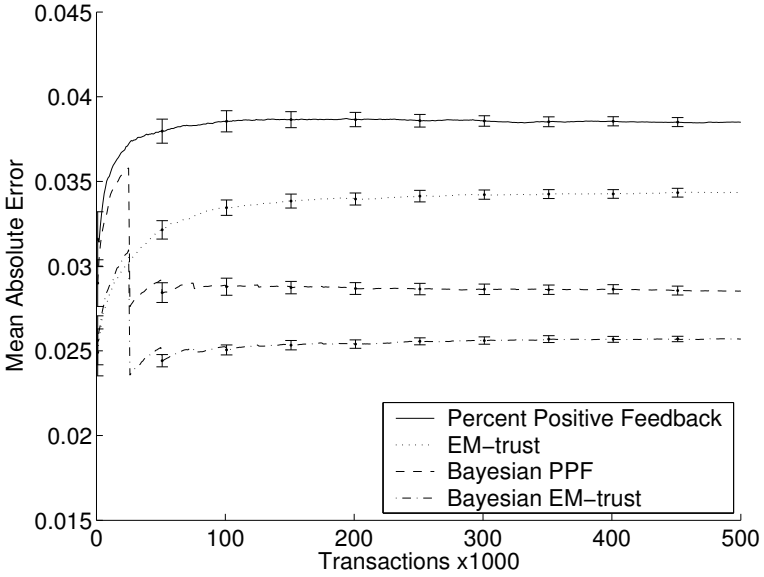


Figure 4.3: Reputation system mean absolute error in Market A with a 50% retaliation rate.

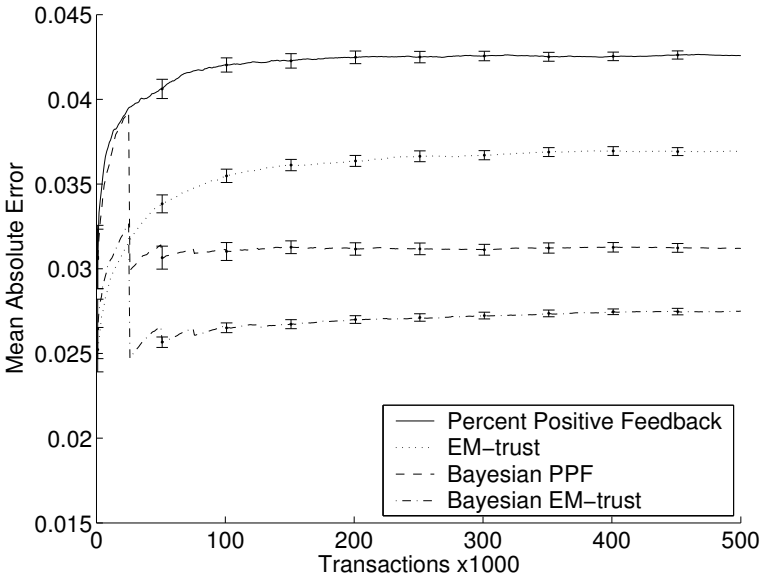


Figure 4.4: Reputation system mean absolute error in Market A with a 100% retaliation rate.

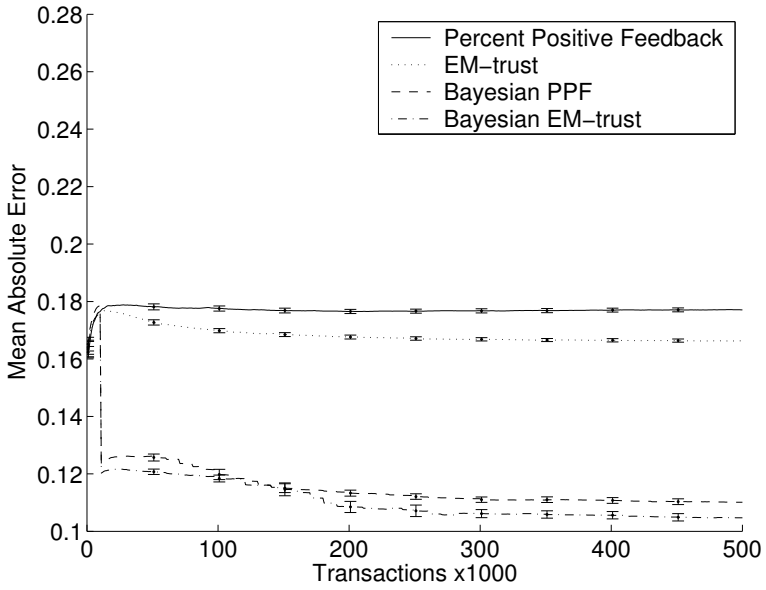


Figure 4.5: Reputation system mean absolute error in Market B with no retaliation.

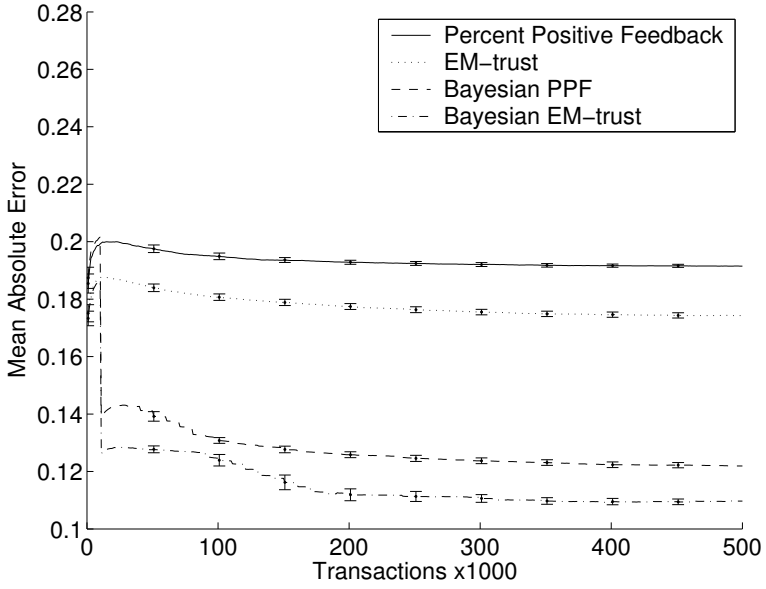


Figure 4.6: Reputation system mean absolute error in Market B with a 50% retaliation rate.

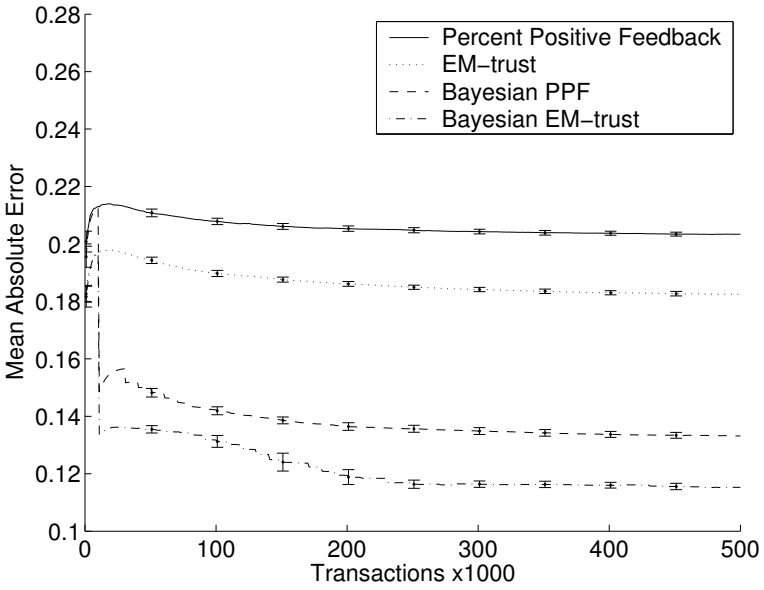


Figure 4.7: Reputation system mean absolute error in Market B with a 100% retaliation rate.

Bayesian algorithms are also less susceptible to errors caused by retaliation than their non-Bayesian counterparts.

4.4.2 Classification Performance

While more accurate evaluations of users' performance is certainly a desirable feature in reputation systems, the fundamental problem they aim to solve is one of classification. A participant in a peer-to-peer marketplace hopes to use the information returned by the reputation system to make a choice about whether to interact or not with a potential trading partner.

To test the algorithms' ability to distinguish good users from bad ones, we look at two statistics: the transaction success rate and the precision of user deactivations. The transaction success rate is simply the percentage of transactions where both parties behaved correctly. The deactivation precision is the percentage of deactivated users (users who leave the system because of a low reputation) whose true reliability is less than the overall mean reliability. In other words, deactivation precision is the percentage of users removed from the system who actually deserve to be removed. The former statistic can be intuitively interpreted as a form of recall and the latter as type of precision. An ideal

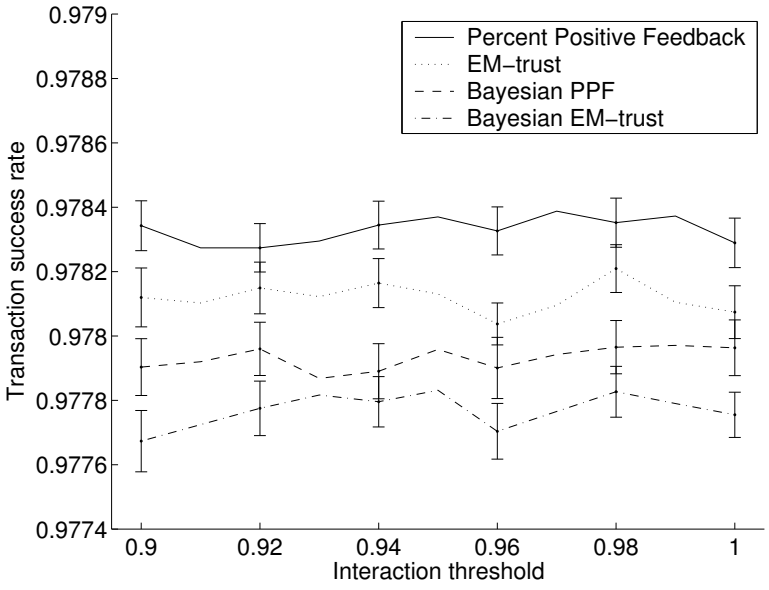


Figure 4.8: Effect of the interaction threshold on the transaction success rate in Market A.

reputation system would score 1.0 on both metrics. We also present an overall performance metric that combines these two results by taking their harmonic mean.

We evaluate the algorithms over a range of interaction thresholds to show how these two statistics change with the selectivity of the marketplace's participants. In Marketplace A, with a mean reliability of 0.9802, we vary the threshold from 0.9 to 1.0. For Marketplace B, with a mean reliability of 0.884, we vary the threshold from 0.76 to 0.96. With thresholds higher than 0.96 in this market, users are so reluctant to interact with all but a tiny subset of their peers that the simulation becomes impractically slow. All tests are conducted with the retaliation rate set to 50% for good users and 100% for bad users. The results are shown in Figures 4.8–4.13.

In both marketplaces, PPF has the highest transaction success rate, followed by EM-trust, then Bayesian PPF, and finally Bayesian EM-trust. By and large, the transaction success rates are not greatly affected by the interaction threshold, with the notable exception of Bayesian PPF, whose performance in Marketplace B rises rapidly at very high threshold values and exceeds EM-trust's performance at 0.96.

That lower error rates should result in lower transaction success rates may seem counterintuitive.

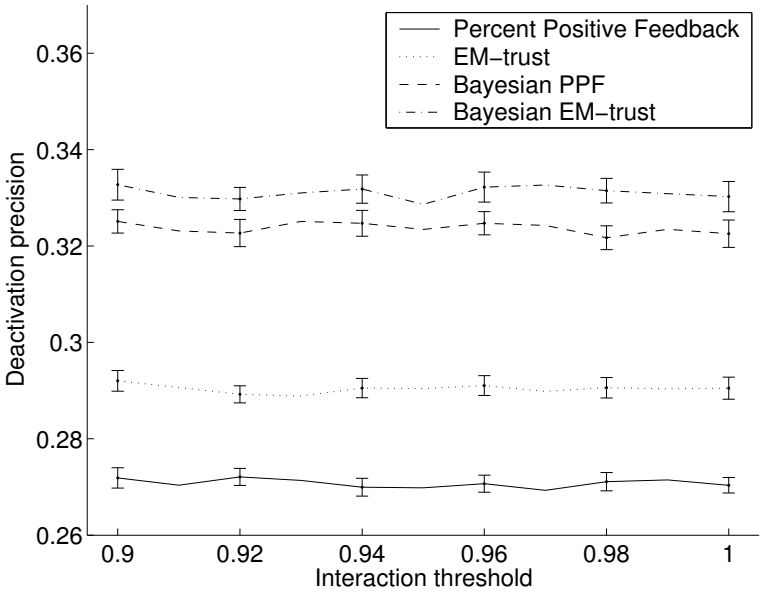


Figure 4.9: Effect of the interaction threshold on the deactivation precision in Market A.

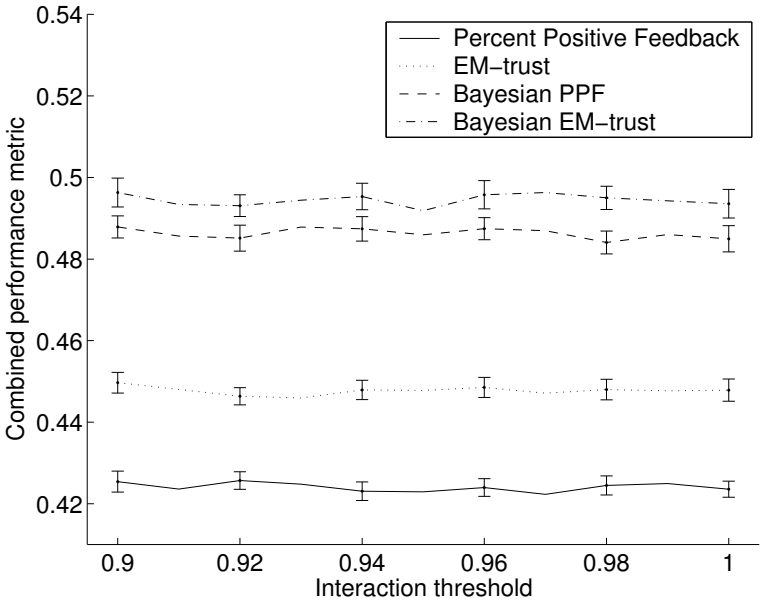


Figure 4.10: Effect of the interaction threshold on the combined precision/recall metric in Market A.

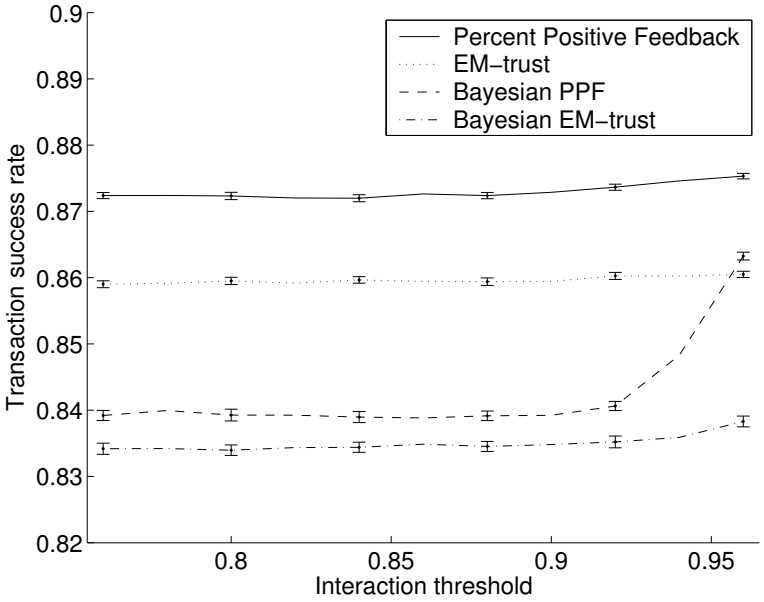


Figure 4.11: Effect of the interaction threshold on the transaction success rate in Market B.

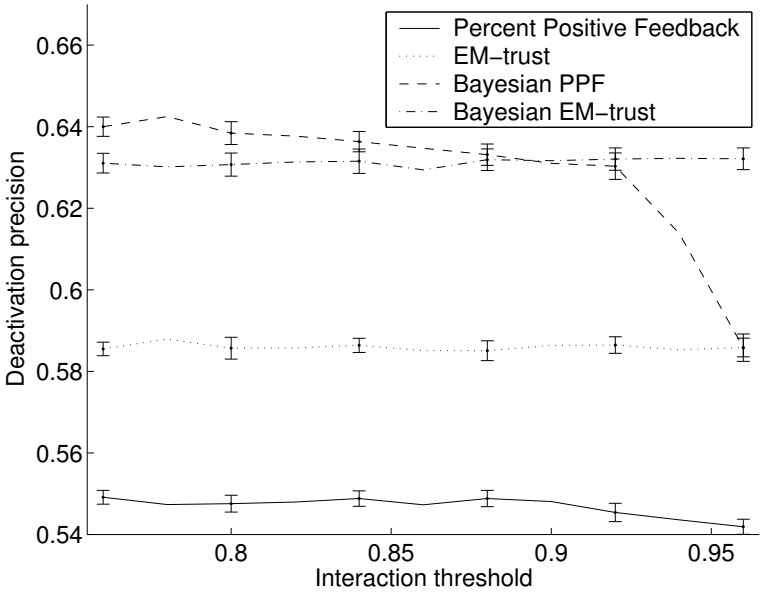


Figure 4.12: Effect of the interaction threshold on the deactivation precision in Market B.

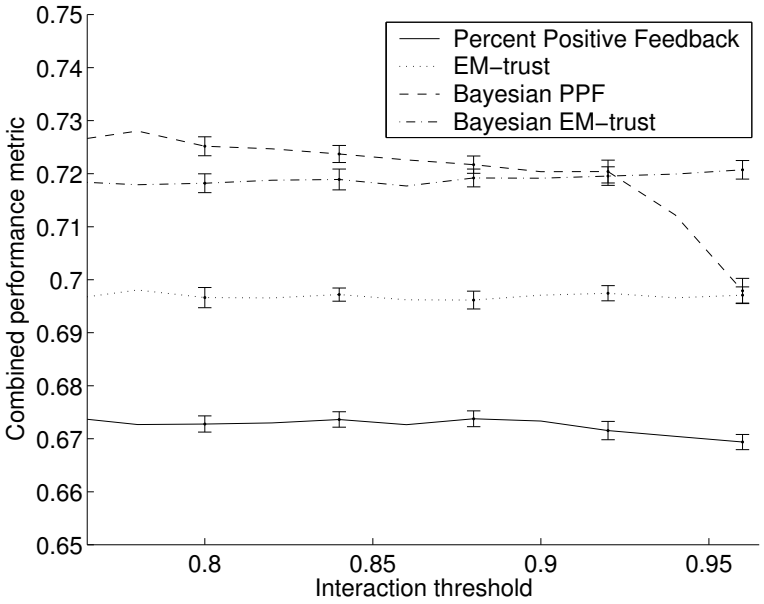


Figure 4.13: Effect of the interaction threshold on the combined precision/recall metric in Market B.

However, we must keep in mind that the MAE metric aggregates both over- and underestimates of users' true performance. PPF, because it does not account for retaliatory negatives, is pessimistic and tends to underestimate true performance. Thus, it effectively detects bad behavior, but at the price of lower estimates for all users. The other algorithms are significantly less pessimistic, but their more accurate estimates fail to detect a small number of the bad transactions found by PPF. However, the differences between the four algorithms on this test are very small: in Marketplace B, the best performance is less than 5% better than the worst performance. In Marketplace A, the algorithms are even closer in performance, with the observed differences being practically insignificant.

In Marketplace A, the interaction threshold has little effect on the transaction success rate. This is likely an artifact of the fact that the distribution of user reliability in this market has a very high mean (0.9802) and very low variance (5×10^{-5}). In such a market, users are likely to be nearly always reliable or nearly always unreliable, so to filter out the bad users, it matters little whether the interaction threshold is set to 90% or 99%.

The interaction threshold does affect all four algorithms to some degree in Marketplace B, with the greatest effect on Bayesian PPF and the least effect on EM-trust. The non-Bayesian algorithms

start to show a slight increase in transaction success around 0.88, while the Bayesian algorithms are nearly unaffected by the interaction threshold until around 0.9-0.91, at which point they start to rise more rapidly.

The deactivation precision reveals greater differences among the four systems in both markets. The algorithms that do better in the transaction success rate test tend to be the worse performers on this test, suggesting that there is a trade-off between precision and recall. In Marketplace A, all of the algorithms show a slight decrease in precision as the interaction threshold increases, but the amount of change is not practically significant. Bayesian EM-trust has the highest precision in this market, followed by Bayesian PPF, then non-Bayesian EM-trust, and finally non-Bayesian PPF.

In Marketplace B, Bayesian PPF performs best at low interaction thresholds, but drops off quickly as the interaction threshold increases. Next comes Bayesian EM-trust, then standard EM-trust, neither of which is strongly affected by the interaction threshold. Finally, non-Bayesian PPF performs the worst, with its precision declining slightly more as the interaction threshold is increased.

The ranking of algorithms and overall trends in the combined metric results look very much like the precision results because of the much greater differences among the algorithms' precisions than among their transaction success rates. If we accept the assumption that both precision and recall are equally important for this task, these results imply that EM-trust and the Bayesian algorithms provide significant improvements to the reputation system precision while giving up only negligible transaction success compared to the baseline PPF system. We also conjecture that this increased precision may in fact have other benefits. Because users are less likely to get unfairly low reputations, they may be more inclined to participate in the reputation system, thus further increasing accuracy.

4.4.3 Market Liquidity

While preventing failed transactions is the obvious first priority of a reputation system, it must be balanced against other marketplace concerns. One of a reputation system's ancillary effects is the influence it has on the liquidity of the market. A reputation system that prevents all but the best

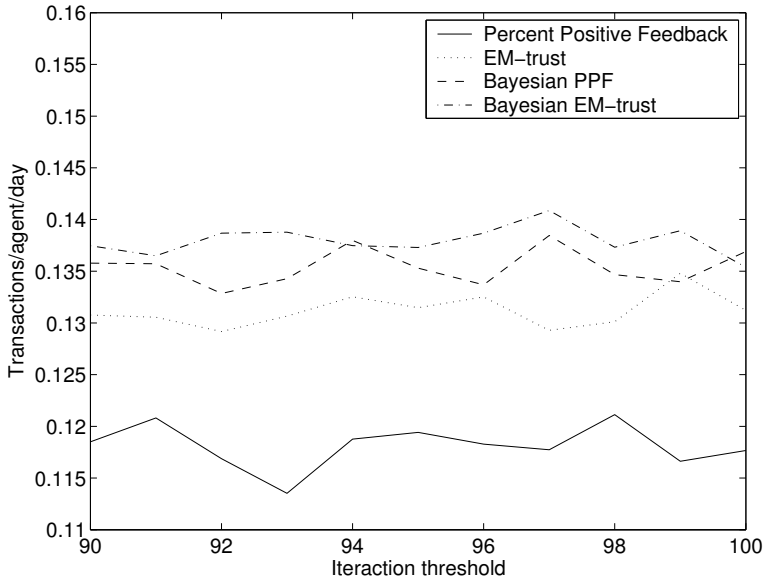


Figure 4.14: Effect of the interaction threshold and reputation system choice on market liquidity in Marketplace A.

users from buying and selling would likely be very safe, but it would also create a market where it is very hard to find a buyer or seller willing to trade.

To measure the effect of the reputation algorithms and the users' interaction thresholds on market liquidity, we look at the average number of transactions per user per simulated day. A higher number of transactions per day indicate a market where it is easier for a user to buy or sell his or her goods. These results are given in Figures 4.14 and 4.15.

In both simulated markets, the liquidity results are similar to the deactivation precision results. This is not surprising, because a market that deactivates fewer good users will likely have more buyers and sellers available at any given time. In Marketplace A, all three of our algorithms have similar liquidity, with the Bayesian systems outperforming standard EM-trust by a small margin. All three are significantly better than the baseline PPF rate.

Once again, there are greater differences among the systems in Marketplace B. Bayesian EM-trust gives the highest liquidity with Bayesian PPF having the next highest at low interaction thresholds. However, the liquidity of the market when using Bayesian PPF drops off dramatically above 0.90 and by 0.96, it has the lowest liquidity of the four. This sudden drop is the main reason we

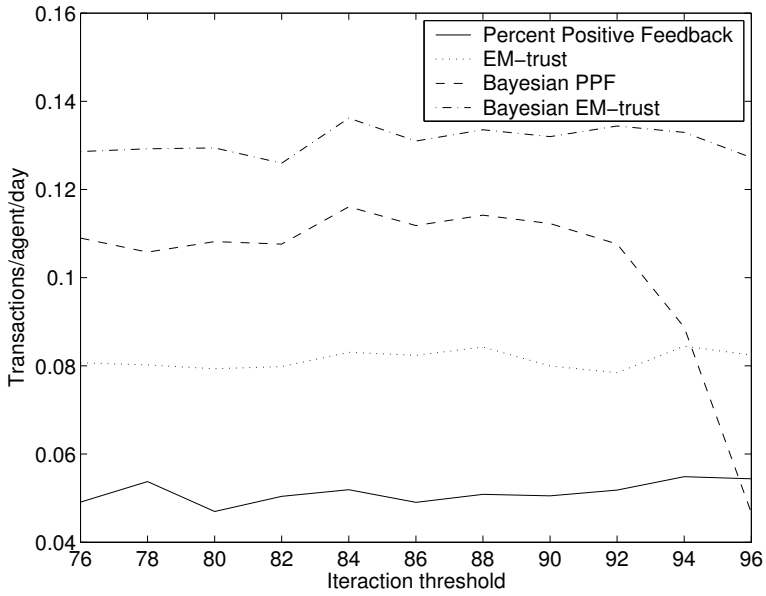


Figure 4.15: Effect of the interaction threshold and reputation system choice on market liquidity in Marketplace B.

are only able to present results through 0.96 for marketplace B: as the liquidity decreases the time needed to run the simulation grows to the point where we cannot finish 500,000 transactions in a reasonable length of time. Standard PPF has the lowest liquidity of the four systems, with standard EM-trust roughly splitting the difference between PPF and Bayesian PPF.

4.5 Discussion

Accurate reputation systems are a requirement for a successful peer-to-peer marketplace. If a reputation system does not provide accurate reputations, it will discourage participation, running the risk of entering a vicious cycle where declining feedback rates cause even less accurate reputations. This problem is particularly serious for reputation systems that give unfairly poor reputations to reliable users, who we most likely expect to provide accurate feedback.

The EM-trust and Bayesian algorithms we present in this report make improvements to the simple eBay averaging approach that we believe will help make reputation systems more accurate and useful. All three of our algorithms estimate true user reliability more accurately than the PPF

algorithm used by eBay. While they permit a negligibly small increase in failed transactions, they give many fewer good users unfairly poor reputations. Additionally, the EM-trust algorithms are far less prone to errors caused by retaliatory negative feedback than their PPF counterparts. Finally, under most circumstances, all three of our algorithms result in a more liquid market, where users are more willing to trade.

In addition to the measurable benefits of these algorithms, we believe that their more accurate reputations will encourage more users to trust and thus participate in the feedback process. We hope that by decreasing the effect of retaliation, we can eliminate some of the hesitation users feel about giving honest negative evaluations. This increased participation will further increase feedback accuracy, creating instead a virtuous cycle that improves both the participation in and the accuracy of the reputation system.

Chapter 5

Sybil-Resitant Reputation Systems

5.1 Introduction

While a serious problem, retaliatory negative feedback is far from the only way in which an attacker can manipulate a reputation system. In this chapter, we examine a different type of attack and discuss several approaches for defending against it.

As discussed in Chapter 1, a major challenge for reputation systems in on-line markets is the fluidity of identity on-line. We must assume that attackers can create new identities at will: the current barriers to account creation, such as a unique email address or a credit card, are relatively trivial to overcome. These new identities can be used to escape bad reputations, which is why we assume that any user with a very low reputation will simply discard his old identity if he has a greater probability of being able to find a partner as a new user.

There is also very little to stop an attacker from creating a large collection of accounts and using them to perpetrate fraud. In a so-called "sybil attack," (30) an attacker uses an army of such fake accounts (the "sybils") to subvert the security of a peer-to-peer system and gain undue influence over it. Not surprisingly, such attacks can be conducted successfully against the reputation system of a a peer-to-peer market.

For example, a seller on eBay can easily create a number of buyer account sybils (only a valid email address is necessary), conduct a fake low value transaction with each (the minimum fee per

transaction is approximately $0.20), and receive a positive feedback from each sybil. Such an attack allows the seller to quickly accrue a long history of positive feedback with minimal expenditure of time and money. Such attacks can be very hard to detect — while dozens of $0.01 sales might attract suspicion from buyers browsing the seller's feedback history, the item description and sale price are deleted from the log after 90 days. Furthermore, with only slightly higher cost (the fee for a $9.99 sale is only $0.92) the attacker can vary the types and costs of items the sybils "buy." The attacker can also space out the "sales" in time, adding a handful of fake transactions for each legitimate one in order to grow his reputation more quickly or adding a burst of fake sales to force a recently received negative further down in the log.

The ease and power of the sybil attack has attracted attention from researchers studying reputation systems and several means of combating the problem have been discussed. One proposed solution is to enforce a one-to-one correspondence between on-line pseudonyms and real identities using a third party service created to guarantee the authenticity of pseudonyms. (45) To date, no such services have been created, and few sites implement any sort of rigorous identity screening when creating an account.

An alternative solution is to use economic effects to control the creation or use of sybils. If we attach a cost to creating user accounts and conducting transactions, it may be possible to render both sybil attacks and fake transactions between real users uneconomical. Bhattacharjee and Goel (9) derive the conditions necessary for a transaction fee to prevent fake feedbacks. It remains unclear, though, whether the fees needed to prevent bad behavior will be low enough so as not to discourage legitimate participation in the system. Using a CAPTCHA[1] (109) or making users perform a computationally expensive task during the account registration and transaction processes can slow down sybil attacks by foiling automated attempts to register or use thousands of accounts. However, these approaches will not stop a determined attacker from registering accounts by hand.

If we cannot stop people from creating sybil users, then the best defense is to detect them, so that we can discount reputation information coming from sybil sources. A recent result (21) proved that any system where reputation is symmetric (i.e. where reputations are invariant under relabeling of nodes) is theoretically vulnerable to sybil attacks. Feldman et al. (41) demonstrate

[1] **C**ompletely **A**utomatic **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part

a scheme that uses maximum flow to form reputations in a simulated file sharing network, which is non-symmetric and effectively resists sybil attacks. Unfortunately, computing maximum flow is expensive: the fastest general algorithm requires $O(nm \log(n^2/m))$ time for each reputation query in a n-vertex, m-edge graph. (55) The amortized constant time approximate algorithm of (41) limits the total number of iterations of the $O(n^3)$ preflow-push algorithm (54), but they present no evidence that this approach will scale effectively to web scale networks.

The EigenTrust system (68) applies the well-known PageRank (93) algorithm to the problem of trust and reputation in peer-to-peer systems. EigenTrust's authors claim it to be resistant to not just sybils but also to collusion by otherwise legitimate users. We show in Section 5.2 that these claims are false and show several mechanisms for using sybils to attack EigenTrust.

We then describe a novel transformation of EigenTrust, Relative Rank, that realizes two important goals. First, it returns reputation metrics suitable for peer-to-peer markets, where both parties need to simultaneously make a decision to interact or not based on the other's reputation. Second, the reputations returned by Relative Rank resist sybil attacks.

Finally, we propose a new algorithm, RAW, that replaces PageRank within the Relative Rank framework. RAW combined with Relative Rank is provably secure against one main class of sybil attack and also provides a strong bound the effectiveness of the other type. Furthermore, RAW is fully personalizable: it can easily return reputations that are specific to the querying user. RAW is thus able to meet the conditions set forward by (21) as a necessary condition for a sybilproof reputation algorithm.

5.2 PageRank as a Reputation System

In order to understand the extensions to PageRank that confer sybil resistance, we must first look at the PageRank algorithm itself. This section serves as a brief summary of PageRank and of EigenTrust, an application of PageRank as a reputation system. For more details on these algorithms, we refer the interested reader to the original PageRank (93) and EigenTrust (68) papers.

5.2.1 The PageRank Algorithm

Let $G = (E, V)$ be a directed graph where every vertex has at least one outgoing edge[2]. Let S, the *start set*, be a vector of length $|V|$ with $||S||_1 = 1$, which defines a distribution across V. Let A be a $|V| \times |V|$ matrix with each element $a_{ij} = 1/|\mathrm{succ}(j)|$ if there is a link from j to i and 0 otherwise, where $\mathrm{succ}(i) = \{j|(i,j) \in E\}$. The matrix A is thus a stochastic matrix that represents the link structure of G.

Define the random walk process $\{X_t\}_{t=1\ldots\infty}$ on G with constant *damping factor* $c \in (0,1)$:

1. $\Pr\{X_0 = i\} = S_i$
2. With probability c, take a step such that $\Pr\{X_{t+1} = i|X_t = j\} = a_{ij}$.
3. Otherwise, restart at a random node: $\Pr\{X_{t+1} = i\} = S_i$.

The process $\{X_t\}_{t=1\ldots\infty}$ is an irreducible, aperiodic, persistent Markov process with a finite state. By the Perron-Frobenius theorem, the process's stationary distribution, R, is the first eigenvector of the matrix $(1-c)S \times \mathbf{1} + cA$, and can be computed with a simple iterative algorithm.

Definition 1. The i^{th} element of R, r_i, is the *rank* or *PageRank score* of node i.

Details of the PageRank algorithm and its applications to web search can be found in (93).

EigenTrust (68) uses PageRank as a reputation system for peer-to-peer file sharing networks. While web links are binary (either a link is present or it is not), trust relationships are described using a range of values, both positive and negative. When constructing the A matrix, EigenTrust therefore uses a more complex normalization procedure. A user i defines his satisfaction with user j, s_{ij} as:

$$s_{ij} = \mathrm{sat}(i,j) - \mathrm{unsat}(i,j)$$

where $\mathrm{sat}(i,j)$ and $\mathrm{unsat}(i,j)$ represent respectively the number of satisfactory and unsatisfactory interactions that user i has had with user j. The elements of the A matrix are defined by:

$$a_{ij} = \frac{\max(s_{ij}, 0)}{\sum_k \max(s_{ik}, 0)}$$

[2]In real networks, some nodes may not have outgoing links. There are several possible solutions to this problem: we could trim out nodes that link to no one, or we have the process restart at a start set node when it hits a dead end (equivalent to adding a link from a dead-end node to all the start set nodes). In our implementation, we do the latter.

Two important consequences of this normalization process are (1) that the random walk now chooses an outgoing link with probability proportional to the user's satisfaction instead of uniformly and (2) that negative satisfaction ratings are essentially discarded: negative trust is treated the same as no trust.

The creators of EigenTrust propose two decision procedures to use when applying this reputation information. In the first procedure, the user always picks the partner who has the highest EigenTrust score. In the second, the user chooses randomly with probability proportional to the potential partners' scores.

5.2.2 Problems with EigenTrust

Despite the optimistic claims in (68), EigenTrust has a number of problems as a reputation algorithm for peer-to-peer markets:

EigenTrust is vulnerable to collusion and sybils. While (68) claim to demonstrate that EigenTrust is robust to collusion, their evaluation is flawed. Consider the simple collusion scenario where a set of users all agree to form a "feedback clique:" they each leave a maximally positive rating for all other members of the clique. Under such an attack, our tests have shown that each member's rank increases. Furthermore, even a single user can construct a network of sybils that will increase his rank as shown in the next section.

EigenTrust does not have a clear decision procedure. In peer-to-peer markets, users need to be able to look at a potential partner's reputation and decide whether to interact or not. EigenTrust scores are more or less a measure of the degree to which a node is "linked in" to the rest of the graph, and this score grows roughly linearly with the number of transactions. Consequently, the decision procedures proposed by (68) are flawed: they tend to select more experienced, but not necessary more trustworthy, partners.

EigenTrust does not use negative feedback. On-line markets often allow both positive and negative feedback. EigenTrust's strategy of discarding this negative information is sub-optimal in this application. Because EigenTrust scores grow linearly with the number of positive links and ignore the negative ones, a user with a fairly high rate of negative feedback can still see unbounded

grown in his EigenTrust score. For example, consider two users, one of whom has received 10 positive feedbacks and no negatives, while the other has received 50 positive and 50 negative ratings. Obviously, we would want the less experienced but 100% honest user to have a better reputation that the one that is honest only half the time. However, under EigenTrust, negative feedback is discarded, so the user with 50 positive ratings nearly always has higher rank than the one with 10.

EigenTrust is vulnerable to attacks by users in the start set. The vertices with positive probability in the start set distribution fill a special role in PageRank-like algorithms. As the starting point for the random walk, these nodes are the source of all authority in the graph. In classical implementations of PageRank, this start set contains all top level domains, weighted uniformly. In EigenTrust, the start set is a set of "trustworthy" nodes established by the management of the reputation system. In both cases, this start set remains the same for all queries, resulting in a symmetric reputation function, which is provably not sybilproof. (21) As we show in later sections, sybil attacks launched by start set members tend to be much more effective than ones started by non members.

Start set nodes also tend to have much higher EigenTrust scores than other nodes, even well-connected ones. In our tests, start set membership confers a "bonus" of approximately 100 positive feedbacks to a node. This elite status is an invitation for corruption: start set members can either exploit their higher reputations directly or, because a link from a high reputation node has a much greater influence than one from a less reputable one, members can offer links to others in exchange for money or something else of value. While the cost of top-level domains (23) and careful selection of trustworthy nodes in EigenTrust can raise the cost and reduce the effectiveness of sybil attacks, they cannot be eliminated. Any choice of a fixed start set leaves us between a rock and a hard place: reducing the number of nodes in the start set limits the set of users that can use start set membership to launch an attack, but also increases the rarity and value of start set nodes, leading to incentives for corruption.

Fortunately, none of these pitfalls is insurmountable. We spend the remainder of this chapter examining these weaknesses and their solutions in detail.

5.3 Sybil Attacks

Broadly speaking, there are two ways in which sybils can be helpful to an attacker: he can use them to increase his own reputation or he can use a sybil, rather than his main identity, to conduct transactions with other users. We first consider attacks designed to increase the attacker's reputation. With PageRank or EigenTrust, if an attacker can alter the random walk process to increase the amount of time it spends at his node, then he can increase his rank. We assume that the only way an attacker can affect the random walk is by engaging in fake transactions with sybils, thus adding links among his main node and the sybils.

Of course, the user can also increase his reputation by engaging in transactions with non-sybil users. So long as these transactions are legitimate, this increase is simply the desirable process whereby a user builds a reputation through honest behavior. Collusion among users remains a problem, but without some means of verifying transaction legitimacy, we believe such collusion cannot in general be detected.

It is also possible to use the sybils to engage in transactions with other users, but this tactic is counter-productive if the attacker's goal is to increase his main node's reputation:

Proposition 1. *Let $G = (E, V)$ be the trust graph excluding the attacker node and all its sybils. Let $G_a = (E_a, V_a)$ be the graph of the attacker node $v_a \in V_a$ and its sybils $\{s_0, \ldots, s_n\} \subset V_a$. Let $G_C = (E_C, V_C)$ be the complete graph with $V_C = V \cup V_a$ and $E_C = E \cup E_a \cup \{(i, j) : i \in V, j \in V_a\}$.*

The rank of the attacker v_a is maximized when all edges (i, j) between nodes in G and nodes in G_a are connected to v_a.

Proof. We present an informal sketch of a proof for this proposition. A more formal proof can be found in (10).

Without loss of generality, assume that V_a is connected.

Consider incoming edges (i, j) where $i \in V$ and $j \in V_a$. If $j = v_a$, then on each transit of (i, j), the random walk will visit v_a, increasing its rank. However, if $j \neq v_a$, then the probability

that the random walk visits v_a after transiting (i, j) is strictly less than one. So, to maximize its rank, an attacker would want to have edges incoming from G to G_a to go to his main node, not one of the sybils.

Outgoing edges (i, j), where $i \in V_a$ and $j \in V$, fall under a similar argument. If $i = v_a$, then all random walks exiting G_a must first visit v_a increasing its rank. If $i \neq v_a$, then it is possible for a random walk to exit G_a without visiting v_a. So to maximize its rank, the attacker should have all outgoing edges connected to v_a. □

If we allow a non-connected V_a, the ideal scenario for the attacker would be to have all edges incoming from G connected to v_a and have all outgoing edges to G connected from a sybil disconnected from all other nodes in V_a. With this topology, a random walk would enter G_a and could never leave, aside from the probability $1 - c$ chance of restarting the walk at a start set node. However, this configuration violates our assumption that all links correspond to feedback left for a transaction: the two nodes in a transaction, not some third node, must leave feedback for each other. Even with a connected sybil network, the amount of time the random walk process spends within the sybil network can be increased by never leaving positive feedback for any non-sybil trading partners. Without any outgoing edges, the only way for the process to exit is via a probability $1 - c$ jump to a start set node.

5.3.1 Attack Types

While Proposition 1 shows that an attacker cannot increase his reputation by cleverly choosing sybils to engage in transactions, it is nevertheless possible to engineer a network of sybils that increases the attacker's score. Informally, a node's EigenTrust score is the ratio of visits to the node to the total number of steps taken by the process, so there are two strategies for increasing it: increase the number of visits to the node or make fewer visits to other nodes.

A *Type I* attack uses sybils to redirect the random walk back at the attacker's node, increasing the number of visits to it. Figure 5.1 illustrates a simple configuration that implements this attack. The attacker creates n sybils and adds both in- and outgoing links between each sybil and the attacker. Provided the attacker has no other outgoing links (or n is much larger than the number of outgoing

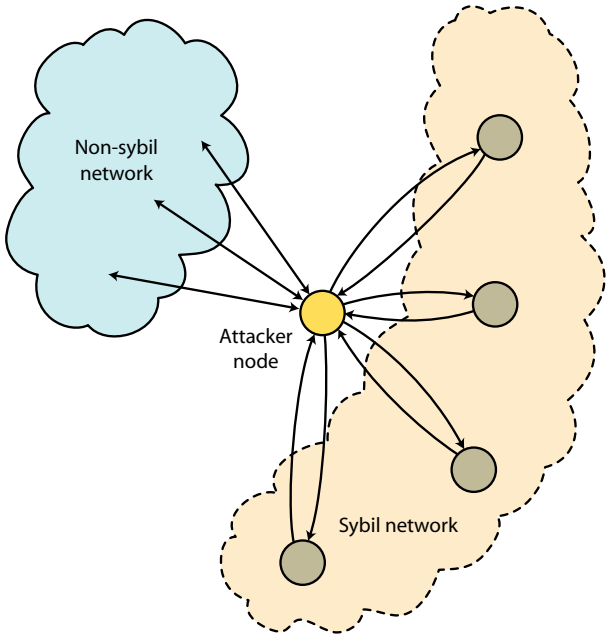


Figure 5.1: Schematic of a Type I sybil attack.

links), once the process enters the sybil network, it will spend approximately half its time visiting the attacker until it resets to a new start set node.

If nodes are not permitted to link to themselves, the configuration in Figure 5.1 yields the largest possible increase in score. The random walk process must spend at least one step away from the attacker before revisiting it, and the illustrated configuration guarantees the process will return to the attack every other step so long as it does not reset to a start set node.

It is possible to bound the effectiveness of the Type I attack in terms of the chosen damping factor. Assume that we have an attacker with the sybil configuration shown in Figure 5.1. Assume the attacker has no outlinks to non-sybil nodes. Once the process is at the attacker's node, the probability of returning to the attacker for exactly n visits is $c^{2n} - c^{2(n+1)}$. Then the expected number of returns to the attacker's node is:

$$\begin{aligned} \sum_{n=1}^{\infty} n(c^{2n} - c^{2(n+1)}) &= \sum_{n=1}^{\infty} c^{2n} \\ &= \frac{c^2}{1-c^2} \end{aligned}$$

The attacker can therefore increase the amount of time spent at his node and thus his score by a factor of $1/(1-c^2)$, which is approximately 3.6 with $c = 0.85$, the choice of damping factor we use for our experiments.

Connections outside of the sybil network will decrease the probability that the process will return to the attacker and will reduce the effectiveness of this attack. However, there may be external reasons (e.g. to dodge suspicion) for an attacker to leave feedback for non-sybils. In such a case, the same bound can be approached in the limit as the number of sybils goes to infinity.

In the *Type II* attack, shown in Figure 5.2, the attacker links to each sybil but does not link back to his main node: each sybil is a dead end. This attack forces the process to restart at a start set node more frequently, preventing visits to nodes outside the sybil network. Sybils are not strictly necessary in this attack: an attacker with no outgoing links at all also achieves the same end. However, if the attacker has outlinks to non-sybil nodes, he will need a significantly larger number of links to dead-end sybils to cause a high restart probability.

The effectiveness of the Type II attack can be bounded to $1/(1-c^3)$ or approximately 2.6 with $c = 0.85$. We prove the bound in Section 5.5.3 for the RAW algorithm, but the same argument holds for PageRank.

While these bounds suggest that both the Type I and Type II attacks are similarly effective, in practice, they have quite different performance. The configurations we use to derive these bounds are both worst-case attacks, but the feasibility of implementing such a worst-case configuration is much higher for the Type I attack. The Type I configuration requires only that the attacker make a sybil network like the one in Figure 5.1 and not leave feedback for any non-sybil user. Both of these requirements are easily met in real markets. The Type II scenario, however, requires very specific configurations of the start set and its members' connections to both the attacker and the rest of the network. Such control over the start set is far beyond the capabilities of the attacker under the present threat model. We believe that a system that permits such control will have far larger problems that the ones discussed here.

As we show in the next section, the Type I attack can be quite effective in practice, while the Type II attack almost never is. However, since these attacks constitute the routes by which

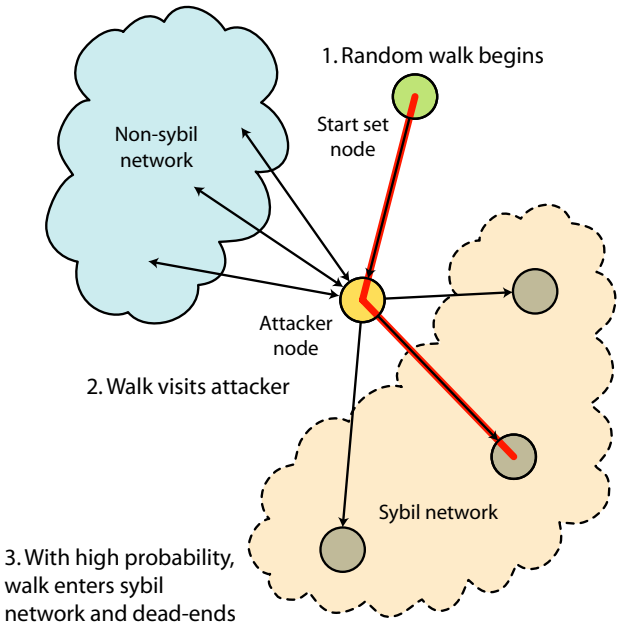


Figure 5.2: Schematic of a Type II sybil attack. The same configuration is used for the Type III attack, but the goal is to create high reputation sybils rather than to increase the attacker node's reputation.

an attacker can potentially manipulate a PageRank-like algorithm to gain a higher reputation, we include them both for completeness.

The *Type III* attack, which can use the same network topology as the Type II attack, has a different goal. Instead of increasing the attacker's reputation, the purpose of this attack is to create sybils with high reputations that can then be spent engaging in uncooperative behavior without affecting the attacker's primary reputation. Once a negative feedback diminishes a sybil's reputation, the attacker simply discards it.

5.3.2 EigenTrust is not Sybilproof

To investigate the effect of these three sybil attacks on the EigenTrust algorithm, we implemented them in our marketplace simulator (described in detail in Appendix A). The simulations are conducted in a market with a bimodal distribution of user honesty: 95% of users behave honestly

Parameter	Value
Proportion of honest users	0.95
Mean honesty of honesty users	0.98
Mean honesty of dishonest users	0.02
Number of buyers	800
Number of sellers	320
Median buy rate (items/day)	0.2
Median sell rate (items/day)	0.5
Market growth rate (% per day)	0.1
Respawn rate	0.667
New user interaction probability	0.05
PageRank damping factor	0.85

Table 5.1: Simulation parameters for the experiments in this chapter.

98% of the time while the remainder behave honestly 2% of the time. Formally, we use a mixture of Betas distribution with parameterization $0.95 \cdot \text{Beta}(98, 2) + 0.05 \cdot \text{Beta}(2, 98)$. Other simulation parameters are given in Table 5.1. We believe that this distribution captures the essence of real networks where users tend to either play by the rules or cheat, and not use some mixed strategy.

Because we do not have an effective decision procedure for EigenTrust, the users in this simulation always interact regardless of reputation.

We measure the effectiveness of the first two attack types by looking at the percentage change in reputation. For the Type III attack, we simply look at the mean reputation of the created sybils. For each test, we ran 10 independent simulations, each with 10 attackers with the final results obtained by taking the mean of all 100 attackers.

Type I attack (Figure 5.5) is clearly effective: even a single sybil causes a measurable increase in reputation and 50 sybils allows the attacker to more than double his reputation. The effectiveness of this attack increases as more sybils are added, but the incremental benefit is less with more sybils. The attack is roughly equally effective whether the attacker belongs to the start set or not; however, the members of the start set begin with much higher reputations, so the absolute increase is greater.

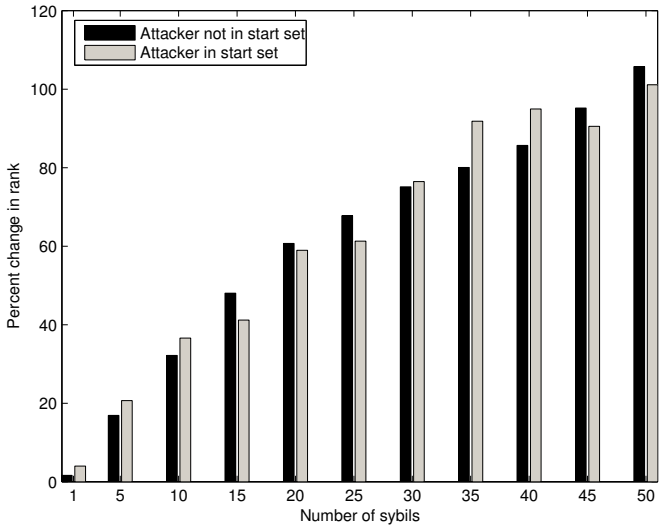


Figure 5.3: Effectiveness of the Type I sybil attack against the EigenTrust reputation system. Performance is expressed as the relative change in the attacker's reputation when sybils are added. All results are the mean of 10 independent trials each with 10 attackers.

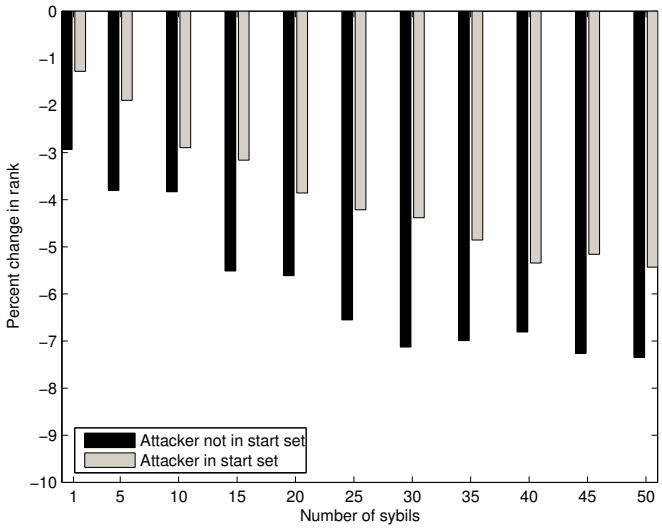


Figure 5.4: Effectiveness of the Type II sybil attack against the EigenTrust reputation system. Performance is expressed as the relative change in the attacker's reputation when sybils are added. All results are the mean of 10 independent trials each with 10 attackers.

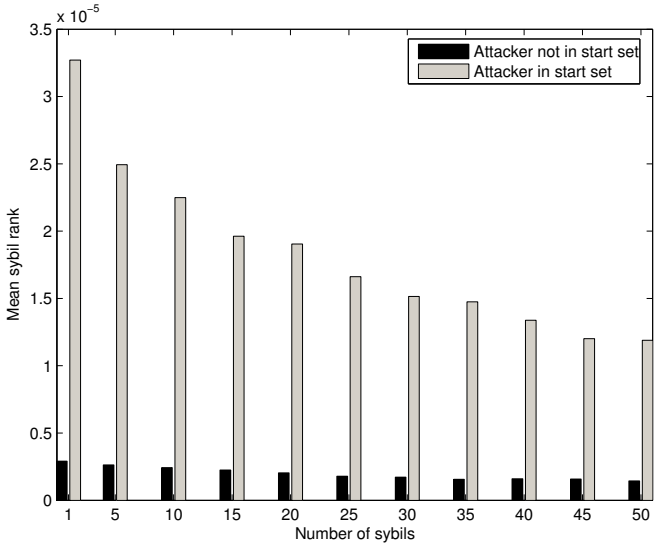


Figure 5.5: Effectiveness of the Type III sybil attack against the EigenTrust reputation system, shown as the average reputation of each sybil. Results are the mean of 10 independent trials each with 10 attackers.

The Type II attack (Figure 5.4) is not effective at all, with sybils causing a decrease in reputation at all levels. It is slightly less ineffective if the attacker is a member of the start set, since the chances of returning to the attacker after restarting a random walk is much higher. While of some theoretical interest, this attack does not appear to be of much concern for practical systems.

It is difficult to evaluate the effectiveness of the third attack (Figure 5.5) because, as discussed in Section 5.2.2, it is unclear exactly what constitutes a *good* or *bad* reputation under EigenTrust. Sybils do receive a positive reputation, though more sybils means each sybil's reputation is slightly lower. More troubling is that the reputations of sybils created by a start set member are, on average, nine times higher than those created by a non-member. Since the configuration of sybils in the Type III attack is identical to that of the Type II attack, we note that a start set member can trade off a small (roughly 5%) decrease in his main identity's reputation in order to create an army of relatively high reputation sybils.

5.4 Relative Rank: PageRank for Markets

We now introduce our technique of *Relative Rank*, a transformation of EigenTrust scores with several desirable properties:

- **Relative Rank has a clear decision procedure.** Honest users, regardless of their experience, receive high Relative Rank scores, while dishonest ones receive low scores, permitting users to base their interaction decision on a simple constant threshold.
- **Relative Rank uses negative feedback.** A user that engages in a steady rate of bad behavior will have a lower Relative Rank than a user whose behavior is consistently honest.
- **Relative Rank resists sybil attacks.** For users that are not members of the start set, Relative Rank does not increase with either Type I or Type II sybil attacks. Furthermore, the sybils created in a Type III attack have reputations too low to reliably engage in transactions on the attacker's behalf.

5.4.1 Relative Rank Defined

The primary motivation for Relative Rank is the transformation of PageRank into a reputation system suitable for use in peer-to-peer markets. In typical markets, potential buyers and sellers examine each others' reputations and try to decide whether or not it is safe to interact. While EigenTrust is proposed as a PageRank-based reputation system, a user's EigenTrust score increases with the number of positive feedbacks received, not with the success rate of the user, making it difficult to discriminate between honest and dishonest users. Additionally, users in the start set begin with much higher rank than non-members making it hard to compare members' and non-members' scores without some understanding of the effects of start set membership. However, enlarging the start set to include all users allows a new, trivial sybil attack[3].

While trivially vulnerable to many attacks including by sybils, Percent Positive Feedback (PPF), the system used by eBay, is a quite good reputation system from a usability standpoint. A PPF score

[3]Essentially, the attacker just creates a large army of sybils and points them at his main node. Because all nodes are in the start set, on each random walk restart, there is a chance that the walk will start at a sybil and proceed directly to the attacker. We do not deal with this attack in depth because it is easy to avoid by not including all nodes in the start set.

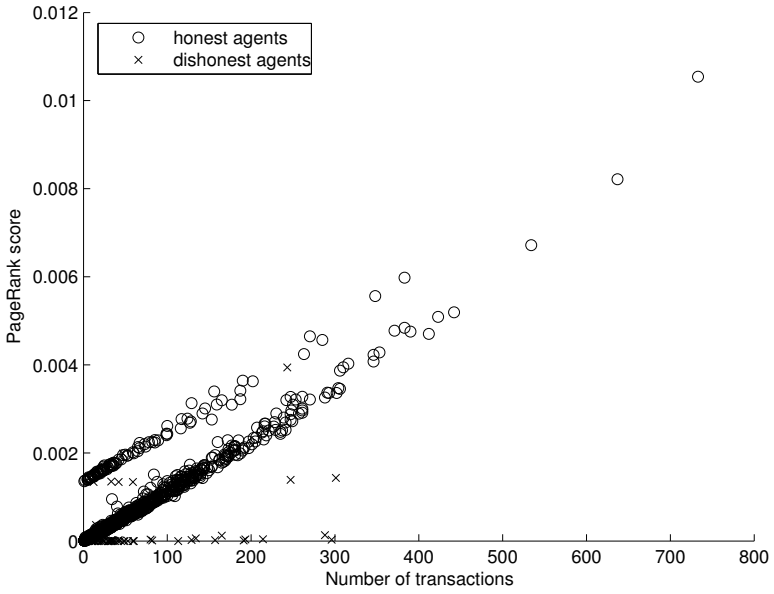


Figure 5.6: EigenTrust score vs. number of transactions for all users in simulated market with a bimodal mixture of Betas distribution of honesty (mean 93.2%).

is essentially just the mean past performance of the user in the market, and the additional count of positive feedbacks received gives some information about the user's experience and thus puts confidence bounds on the performance estimate. Ideally, we would like a system with reputations as easy to interpret as PPF's but without its many vulnerabilities.

Figure 5.6 plots EigenTrust score against the number of transactions for all users in two simulated markets. In the first market, we use the bimodal distribution of agent described in Section 5.3.2. In the second market, user honesty is distributed uniformly. We use this second fairly unrealistic distribution primarily to illustrate how PageRank handles different levels of honesty.

Examining Figure 5.6, we see four major regimes:

1. Honest agents whose ranks follow a line with positive slope and intercept 0.0015
2. Honest agents whose ranks follow a line with positive slope and intercept 0
3. Dishonest agents whose ranks lie around 0.0015, regardless of experience
4. Dishonest agents whose ranks lie around 0, regardless of experience

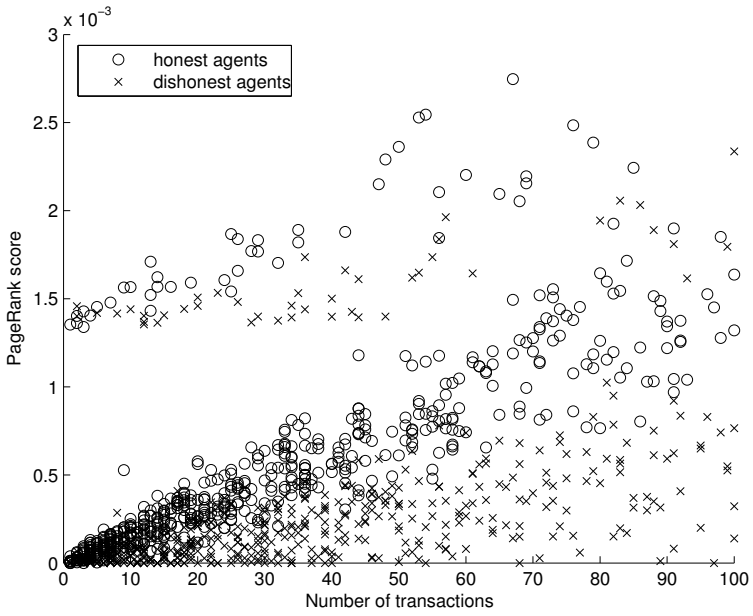


Figure 5.7: EigenTrust score vs. number of feedbacks received for all users in a simulated market with uniformly distributed honesty (mean 50%).

Encouragingly, the rank of dishonest agents behaves differently than that of honest ones. However, it is clear that a simple threshold will not work very well: a threshold less than 0.0015 will miss many dishonest users, while one much greater than 0 will classify a large number of honest agents incorrectly. Groups 1 and 3 represent users that belong to the start set and the other groups consist of non-members. However, even if we divide the users based on start set membership, any threshold we set will likely exclude a large portion of users with low experience.

Similar patterns are apparent In the uniformly honest market (Figure 5.7). Once again, start set members receive a bonus of about 0.0015 regardless of their honesty. The users' ranks are distributed across a triangular region in the first quadrant, with a roughly linear boundary between honest and dishonest ones. If we plot only users of a fixed level of honesty, the plotted points roughly follow a ray beginning at the origin (or at $(0, 0.0015)$ for start set members) and extending into the first quadrant. The angle this ray forms with the x axis is proportional the user's honesty.

This observation forms intuition behind the Relative Rank algorithm:

1. Run EigenTrust.

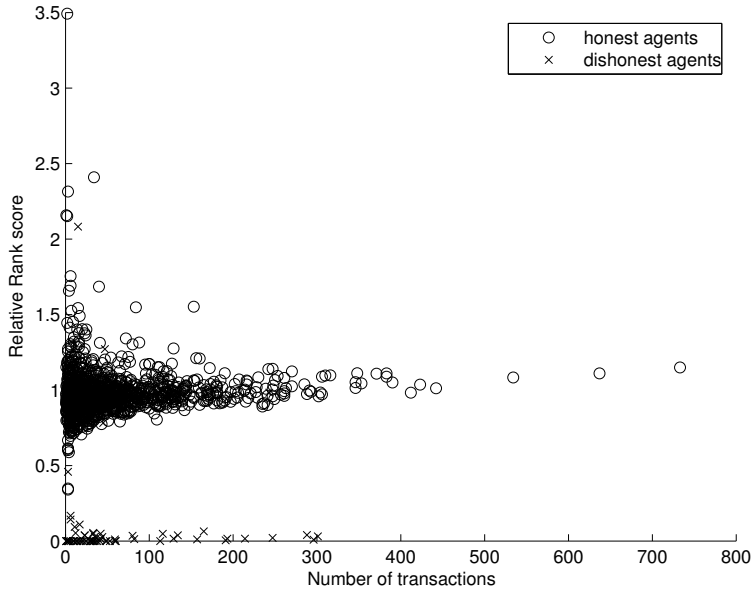


Figure 5.8: Relative Rank versus number of transactions for a markets with bi-modal honesty distribution (mean 93.2%).

2. Separate start set members from other users.

3. For each k, find the non-start-set user i_k that has the highest observed rank, r_{i_k} among users who have received k feedbacks, including *both* positive and negative feedback.

4. Fit a line to the pairs (k, r_{i_k}) and obtain a slope, $\beta_{\bar{S}}$, and intercept, $\alpha_{\bar{S}}$.

5. Repeat steps 3 and 4 for start set members to obtain a separate intercept and slope, α_S and β_S.

For a non-start-set user i with k feedbacks, define the *Relative Rank score* as:

$$s_i = \frac{r_i - \alpha_{\bar{S}}}{\beta_{\bar{S}} k}$$

The same definition holds for start set members, except that α_S and β_S are used.

Similar plots of Relative Rank versus number of transactions for the two example markets can be found in Figure 5.8 and 5.9. Clearly, a simple linear separation between honest and dishonest users appears to be a reasonable approach in both of these markets.

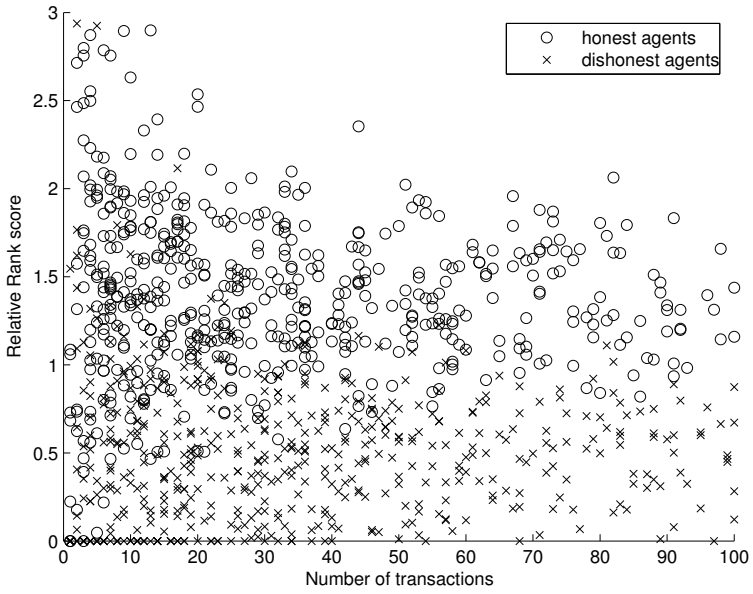


Figure 5.9: Relative Rank versus number of transactions for a markets with uniformly distributed honesty (mean 50%).

5.4.2 Reputation System Performance

Before we look at its performance with sybils, we examine how well Relative Rank serves as a reputation metric. Certainly, the ability to resist sybils is moot if the system cannot sort out good users from bad.

Figure 5.10 presents a ROC curve that illustrates the trade-off between detecting dishonest users and incorrectly labeling honest users as dishonest when using Relative Rank with a simple fixed threshold in the two example markets. The area under this curve is considered a good non-parametric estimation of a classification algorithm's performance, with an ideal system having area 1. For Relative Rank, the area under the curve is 0.9306 for the market with uniform honesty and 0.9212 for the market with a bimodal honesty distribution. In both cases, we define an honest user as one whose rate of successful transactions is equal or greater to the mean. If we relax this definition somewhat so that an honest user is one that behaves correctly 90% of the time, the area under the curve for the bimodal market increases to 0.996.

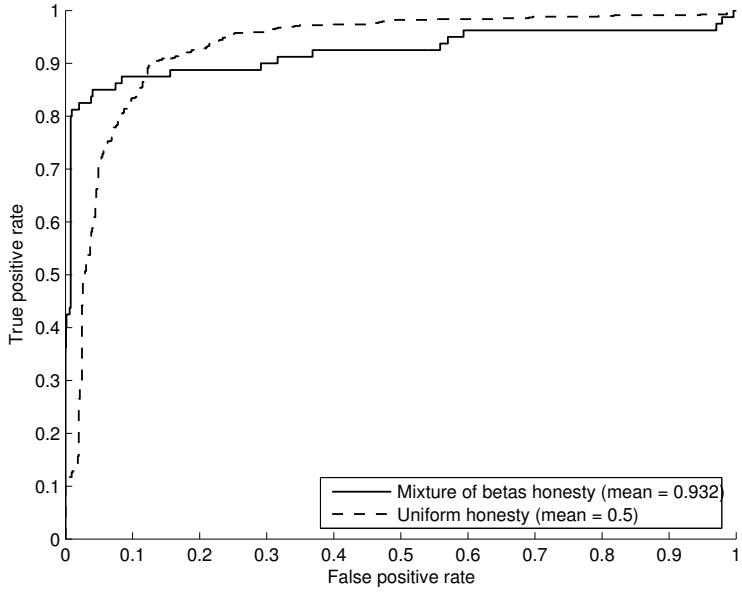


Figure 5.10: ROC curve illustrating the true positive/false positive trade-off when using Relative Rank to identify dishonest users in the two simulated markets.

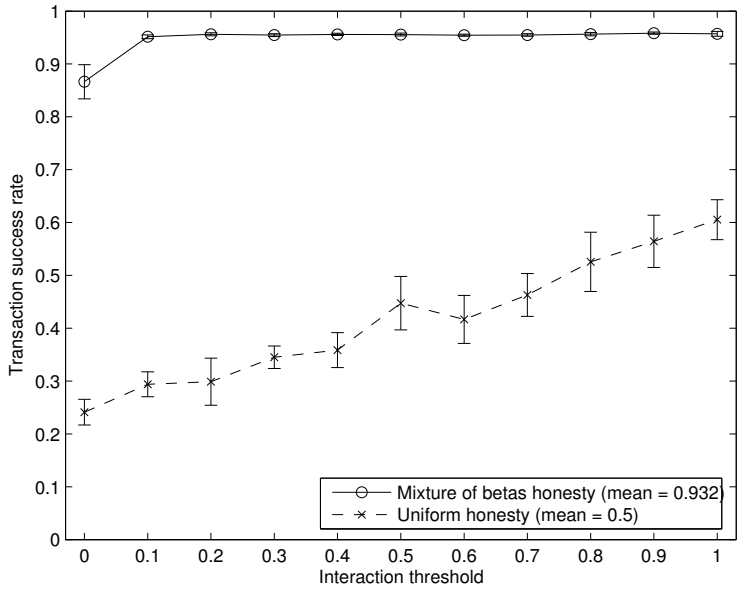


Figure 5.11: Effect of the interaction threshold on the transaction success rate with the Relative Rank reputation system in the two simulated markets. Results are the mean of 10 indepedent runs; error bars indicate standard deviation.

In Figure 5.11, we measure the transaction success rate (the percentage of transactions where both parties behave honestly) in the example markets. We examine the markets' performance at several different interaction thresholds (the minimum reputation an agent must have before being allowed to interact).

In the bimodal market, Relative Rank nearly perfectly separates the good and bad users. With a threshold of 0 (all users always interact) the observed transaction success rate is 0.866, very close to the rate of 0.869 expected without a reputation system. However, with Relative Rank and a moderate positive threshold (0.4–0.6), the success rate increases to 0.956, just slightly less than the 0.960 rate expected if only the good users are permitted to operate. However, Relative Rank seems less capable of discriminating between fine differences in agent honesty: increasing the threshold beyond a moderate level does not provide a significant benefit. This is not unexpected: with roughly equal honesty and experience, there will be some variation in users' Relative Rank scores depending on the local topology of the graph in which they operate. We do not view this as a problem — there is ample evidence that suggests that a bimodal distribution of users with a mostly honest majority and a dishonest minority is a reasonable model of real user behavior. Furthermore, it is exactly this sensitivity to graph structure that gives Relative Rank its resistance to sybil attacks.

The transaction success rate in the market with uniform honesty is much lower overall because even if only users with above average (> 0.5) honesty are permitted to operate, the expected rate of transaction success is only 0.563. Even under these difficult conditions, Relative Rank with a high (> 0.8) threshold is able to exceed this target and roughly halves the number of failed transactions.

5.4.3 Relative Rank and Sybils

With Relative Rank established as a useful reputation algorithm for peer-to-peer markets, we examine its behavior under the three sybil attack scenarios described in Section 5.3.1. All of these experiments are run only in the market with the bimodal user honesty distribution. The results of this experiment are shown in Figures 5.12–5.14. Comparing these graphs with the results for EigenTrust (Figures 5.3–5.5), we see that Relative Rank is significantly more resistant to sybil attacks.

The Type I attack (Figure 5.12) is now completely ineffective for users that do not belong to that

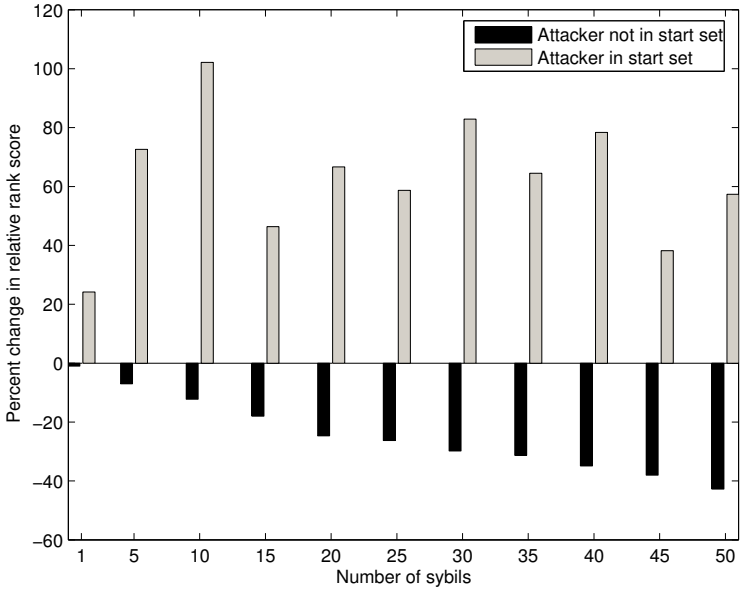


Figure 5.12: Performance of Relative Rank under the Type I sybil attack scenario described in Section 5.3.1. Results are the mean of 10 independent simulations with 10 attackers in each trial.

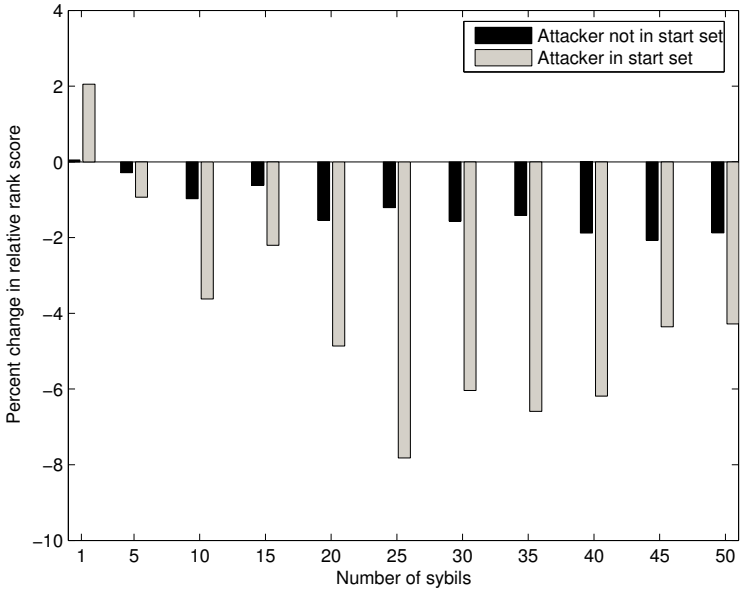


Figure 5.13: Performance of Relative Rank under the Type II sybil attack. Results are the mean of 10 independent simulations with 10 attackers in each trial.

start set but remains a viable means for start set members to increase their reputations. Non start-set members see their Relative Rank scores decline further with each sybil added. Start-set members can increase their reputation by about 50% with sybils, but there is no indications that adding more than a few sybils offers any benefit.

While we cannot prove that Relative Rank is completely resistant to Type I attacks by non-start-set members, we do have an intuitive explanation for this observed behavior. Each inlink added to a node potentially increases the chances that the random walk process will visit the node. A connection to a high rank, highly-connected node will result in a greater rank increase than one to a non-connected node. The slope fit by the Relative Rank algorithm essentially measures the average increase in rank for each positive connection. Because the sybils are only connected to the trust graph via the attacker, their connection offers a smaller than average increase in rank and therefore causes a decline in Relative Rank.

The Type II attack (Figure 5.13) is, once again, more or less useless: nearly all attackers see their Relative Rank fall with sybils. One exception is for start set nodes with only one sybil, which gives a very small reputation increase, but this small increase is of little practical benefit to the attacker.

Since, unlike EigenTrust, we have an interaction decision procedure for Relative Rank, we can analyze the impact of the Type III attack (Figure 5.14) more thoroughly. The results of the previous section suggest that a good interaction threshold for this example market is around 0.5. All of the sybils created by non-start set users are thus useless: their reputation is below the interaction threshold, so it is unlikely that the attacker can use them to engage in any transactions.

However, sybils created by start set members have very high reputations. An attacker can thus create a large number of sybils with only minimal effect on his main identity's reputation and conduct a large number of fraudulent transactions before the sybils' reputations are expended. If used to commit fraudulent transactions, f negative feedbacks will reduce a sybil's Relative Rank by a factor of $1/f$, so with 25 sybils, each sybil can be used until it accumulates approximately 3 negative feedbacks.

While initially envisioned as merely a way of adapting EigenTrust to peer-to-peer markets,

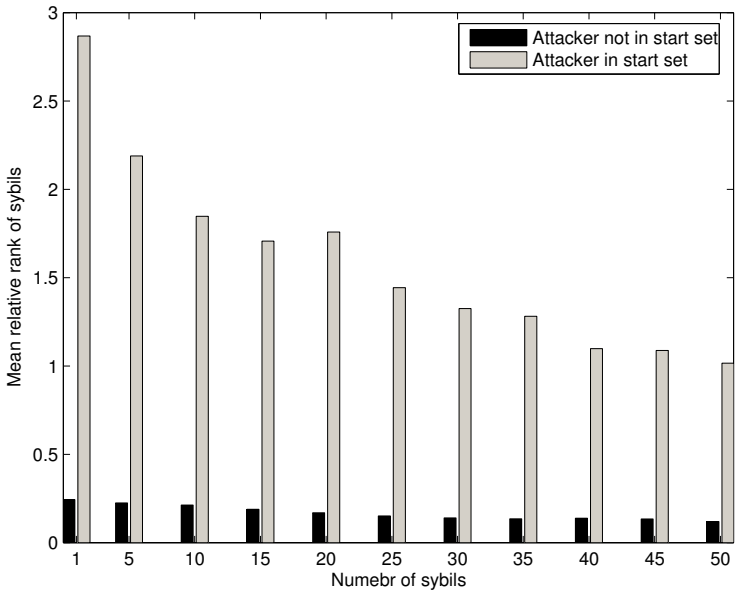


Figure 5.14: Performance of Relative Rank under the Type III attack.

Relative Rank has unexpected benefit of increased resistance to sybil attacks, at least by attackers that do not belong to the start set. However, it is still vulnerable to abuse by start set members. We also cannot prove this sybil resistance: it appears to be generally true, but may simply be an artifact of our choice of simulation parameters.

5.5 The RAW Algorithm

To address the few remaining concerns with Relative Rank, we introduce RAW, a PageRank-like algorithm with two important properties:

1. Provable immunity to Type I attacks and a provable bound on the effectiveness of Type II sybil attacks.

2. Asymmetric, personalized reputations, which render attacks that rely on start set membership ineffective.

RAW does not replace Relative Rank; rather, it replaces the PageRank algorithm within the core of the Relative Rank framework. The combination of RAW with Relative Rank achieves our goal of a provably sybil resistant reputation system for peer-to-peer markets.

5.5.1 Definition of the RAW Algorithm

The setup for RAW is the same as for PageRank: we have a directed graph, $G = (E, V)$, representing the users and their trust relations as well as a start set, S and constant damping factor, $c \in (0, 1)$. The RAW process, $\{(X_t, H_t)\}_{t=1...\infty}$ is a random walk on the graph that proceeds according to the following rules:

1. $H_0 = \emptyset$, $\Pr\{X_0 = i\} = S_i$.
2. With probability c, set $H_{t+1} = H_t \cup \{X_t\}$ and take a step such that $\Pr\{X_{t+1} = i | i \in H_t\} = 0$ and $\Pr\{X_{t+1} = i | X_t = j, i \notin H_t\} = a_{ij} / \sum_{k \in \mathrm{succ}(j) \setminus H_{t+1}} a_{kj}$.
3. Otherwise, $H_{t+1} = \emptyset$ and $\Pr\{X_{t+1} = i\} = S_i$.

Definition 2. If R is the length $|V|$ vector describing the stationary distribution of X_t in the process $\{(X_t, H_t)\}_{t=1...\infty}$ defined above, then R_i is the *RAW score* of node i.

This process is very similar to the one used to define PageRank with one important difference: the process cannot visit the same node more than once between resets to a random start set node. This property is the key to RAW's sybil resistance. No configuration of edges can cause the process to revisit a node, so the Type I attack is impossible by definition.

RAW behaves very similarly to PageRank in the absence of sybils and can be used as a "drop-in" replacement in EigenTrust, Relative Rank, or any other system that uses PageRank.

5.5.2 Implementation and Personalization

The addition of history obviously renders the RAW process non-Markov, so simple closed-form or iterative formulations of its stationary distribution are not readily apparent. For the experiments in this paper, we use a Monte Carlo implementation that directly simulates the random walk process.

For deployment in a web-scale marketplace, it will be necessary to efficiently scale up this implementation from thousands to millions of nodes. Similar techniques have been proposed for Personalized PageRank web search systems (44), and these systems can be readily adapted to computing RAW scores instead.

A key benefit of this implementation of RAW is that it can be fully personalized. To accomplish this, we create a "meta-start-set," a collection of start sets, each with only a single member. We then run the Monte Carlo simulation once for each element of the meta-start-set to build a "fingerprint" of ranks that result when just that single node is used as the start set. These fingerprints are stored in a database for easy access.

At query time, the user constructs a start set by forming a union of several meta-start-set elements, looks up the RAW score of the target in the fingerprints of each start set member, then constructs a personalized RAW score by taking the (optionally weighted) average of the queried fingerprint values. In this way, the user creates the start set dynamically for each query. A proposition in (44) proves that a start set built up in this fashion is equivalent to a start set chosen in the standard way.

In a practical system, the market administration will want to use a meta-start-set large enough to offer a user a wide choice of possible start set nodes, yet small enough to make the Monte Carlo RAW calculation tractable. Provided the meta-start set is large enough, a user will be able to find a sufficiently large start set that does not include either the node whose reputation is being queried or any of its immediate neighbors, drastically reducing the effectiveness of sybil attacks that rely on start set membership or proximity.

5.5.3 RAW and Sybils

The proof of RAW's immunity to Type I attacks is by definition: RAW prohibits multiple visits to the same vertex between resets to a start set node, so any configuration of sybils that attempts such a redirection will fail. Obviously, this immunity to Type I attacks also carries over to RAW Relative Rank: feedback from sybils cannot increase the RAW score, but it does increase the feedback count, thus decreasing Relative Rank score.

Type II attacks are theoretically possible against RAW; however, we can prove a tight bound on their effectiveness.

Proposition 2. *Let r_i be the RAW rank of a user, i, without any sybils and let r_i' be the RAW rank of the same user after creating sybils in a Type II configuration. If c is the chosen damping factor, then the effectiveness of the attack is bounded by*

$$\mathbb{E}\left[\frac{r'}{r}\right] < \frac{1}{1-c^3}$$

Proof. We consider the worst case for an attacker that is not in the start set[4]: there is a single start set node, s, that is the source of all random walks. It is connected directly to i and to no other nodes. This configuration maximizes the number of visits to i, because i lies along the path of all walks of length 2 or more. The attacker has connections to n non-sybil nodes.

The expected number of visits to i on each walk is simply the damping factor c. The expected walk length given a visit to i is $1 + c + c^2(1 + l)$, where l is the expected length of a random walk in the non-sybil portion of the graph. So, the expected rank of i without sybils is:

$$\mathbb{E}[r] = \frac{c}{1 + c + c^2(1 + l)}$$

When i creates m sybils in a type II configuration, the walk transitions from i to a sybil with probability $m/(m + n)$, so the expected rank with sybils is:

$$\mathbb{E}[r'] = \frac{c}{1 + c + c^2(1 + \frac{m}{m+n}l)}$$

If we take the limit as $m \to \infty$, we get that:

$$\mathbb{E}[\frac{r'}{r}] = \frac{1 + c + c^2(1 + l)}{1 + c + c^2}$$

If the random walk never hits a dead end, then $\mathbb{E}[l] = c/(1 - c)$. Because dead ends are possible, $\mathbb{E}[l]$ is strictly less than this value. Making this substitution for l gives us our bound. □

[4]For an attacker in the start set, the worst case is when the attacker is the only start set node. In this case, the attacker can capture *all* the random walk on each step, giving a score of 1 to the attacker's node and 0 to all others. We consider this case somewhat degenerate, as any system that permits the attacker this much control over the start set is essentially helpless.

For the choice of $c = 0.85$ used in our experiments, the maximum increase in reputation with an attack of this type is approximately 2.6. We can also solve the above equation for c, given a desired bound on r'/r.

In practice, attacks of this form are far less effective than this bound suggests. There are typically many start set nodes, most of which do not connect to the attacker's node, making the probability of returning to the attacker extremely low. Those that do connect to the attacker will typically connect to other nodes as well, further diluting the effectiveness of this attack. Nodes in the start set have the most likely chance of benefiting from the Type II attack, but in all of our experiments, they are not able to make effective use of this attack. Personalization blunts this attack's potential even more, because the attacker can no longer count on start set membership or a favorable relationship with a start set node.

5.5.4 Results

Figure 5.15 plots the transaction success rate against the interaction threshold for RAW Relative Rank in our simulated market. Compared to standard Relative Rank (Figure 5.11), there are few differences. Both systems are about equally effective at preventing failing transactions. However, the RAW version experiences a slight reduction in transaction success with high (> 0.8) interaction thresholds, due to higher score variances introduced by the Monte Carlo implementation. Once again, a moderate interaction threshold of around 0.5–0.7 makes the best trade-off between preventing failed transactions and not deactivating too many honest agents.

Performance with sybils (Figure 5.16) is as predicted by theory. Neither Type I nor Type II sybil attacks achieve any practical measure of success in increasing the attacker's RAW Relative Rank. Sybils created in a Type III attack (Figure 5.17 have RAW relative ranks in the 0.25–0.45 range, similar to what we saw with standard Relative Rank for non-start set members. However, with RAW Relative Rank, the "start set" disappears as a concept, so it is not possible for an attacker to exploit his start set membership to launch a successful Type III attack.

Overall, RAW Relative Rank achieves all our goals: it is an effective reputation algorithm for

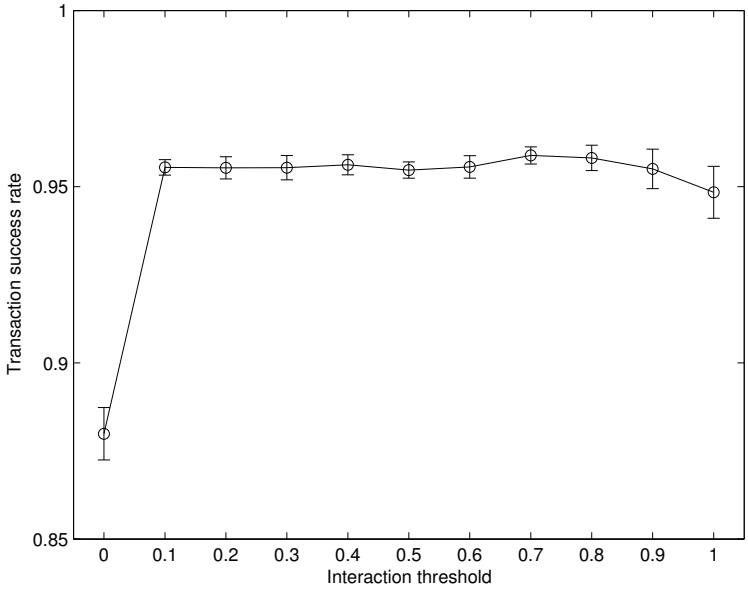


Figure 5.15: Transaction success vs. interaction threshold for the two simulated markets using the RAW Relative Rank reputation system. All results are the mean of 10 independent trials; error bars indicated standard deviation.

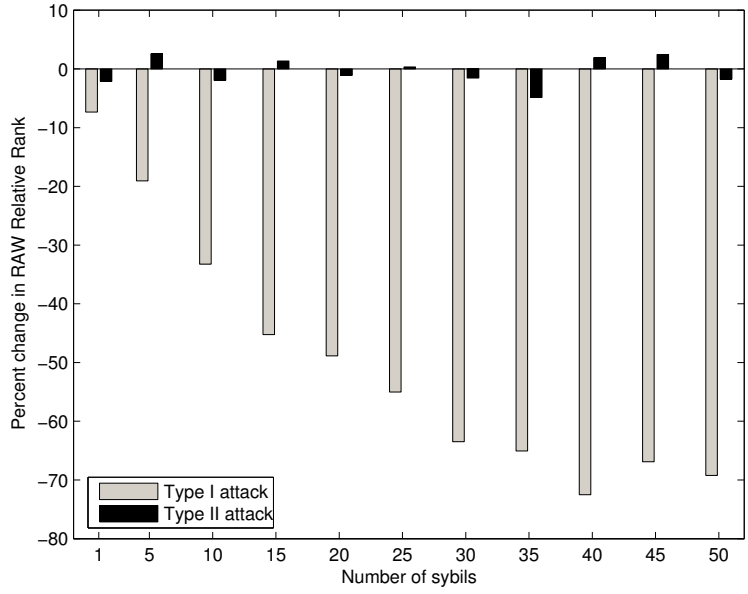


Figure 5.16: Performance of RAW Relative Rank against Type I and II sybil attacks in the simulated bi-modal market. Results are the means of 10 independent trials each with 10 attackers.

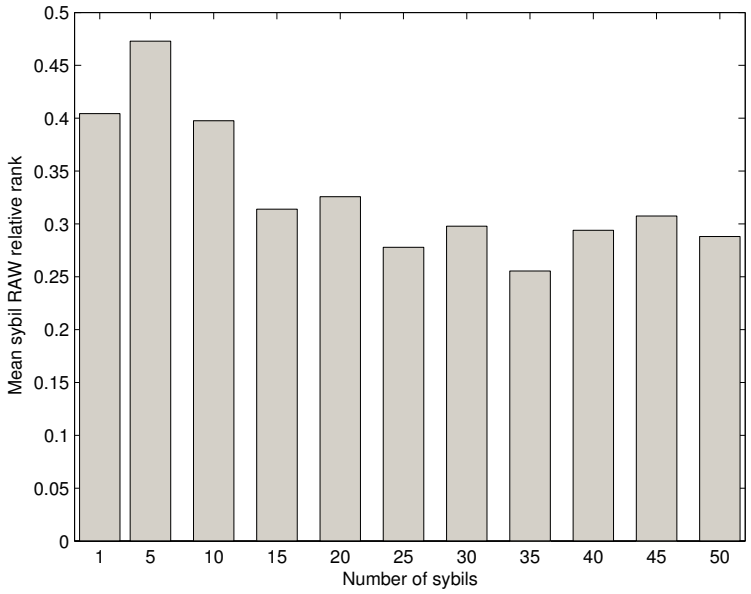


Figure 5.17: Effectiveness of Type III sybil attack against RAW Relative Rank in the simulated bi-modal market. Results are the mean of 10 independent trials each with 10 attackers.

peer-to-peer markets, it is highly resistant to sybil attacks, and it eliminates the corrupting influence of the start set.

5.6 Discussion

In this chapter, we present two techniques that make considerable progress towards the goal of a fully robust reputation system for peer-to-peer markets. The Relative Rank algorithm takes the widely studied PageRank algorithm and adapts it for use as a marketplace reputation system. It transforms users' EigenTrust scores, which are dependent on their experience level, into a reputation metric that can be easily thresholded against for making trust decisions. Furthermore, it incorporates negative feedback so that users must maintain a high degree of honesty in order to be judged worthy of interacting. Finally, Relative Rank is more resistant to sybil attacks than PageRank: for non-start set users, all three of the sybil attacks we identified fail.

The RAW algorithm replaces PageRank within the Relative Rank framework resulting in several

key benefits. Unlike PageRank, RAW is, by definition, invulnerable to Type I sybil attacks. Type II attack success can be bounded, and in practice is far lower than even the bound suggests. The sybils produced in a Type III attack get reputations that are too low to be useful.

Finally, RAW is completely personalized: the querier can choose the start set, so reputations are asymmetric. Personalization also gives the user the opportunity to receive reputation scores that are based on individual perceptions about which users are trusted sources in the network, rather than forcing them to accept a externally imposed choice. Combined with Relative Rank, RAW becomes a reputation algorithm with a simple decision procedure for peer-to-peer markets, resistance to all three classes of sybil attacks, and no opportunity for corruption by start set members.

Chapter 6

Conclusion

By facilitating trust between otherwise unfamiliar partners, reputation systems help enable peer-to-peer markets' growth. While direct trading between individuals is as old as commerce itself, the adoption of cheap, global communication networks has catalyzed these markets' transformation from local flea markets and classified pages into a significant new means of buying and selling. The social and economic impact of on-line peer-to-peer markets is indisputable: in 2005, nearly $45 billion in merchandise was sold through eBay alone, over twice the revenue of Federated Department Stores, a major traditional retailer. (36; 40)

However, in traditional, local peer-to-peer markets, buyers can examine merchandise before purchasing, and sellers can demand payment in cash, minimizing risks for both parties. On-line, sellers still have protection — they can insist on cash-like instruments or wait until payments clear — but buyers are left trying to judge item quality or condition from low resolution photos and textual description written by sellers with unverifiable credibility. Not surprisingly, such circumstances result in a market where both adverse selection and moral hazard play significant roles. In theory, reputation can mitigate both of these effects by allowing buyers to discriminate between sellers who describe their products accurately and offer good service. In practice, as seen in the studies discussed in Chapter 2, reputation systems often appear to have a positive effect, although the magnitude of their contribution is still a matter of some debate.

Unfortunately, the very features that make reputations useful also provide incentives for abuse.

A high reputation has been observed to attract more bidders and may allow a seller to charge a price premium. Ideally, a good reputation's value will encourage users to accumulate positive feedback through honest behavior. However, it also encourages the search for expedient alternatives for building a reputation. Reputation systems, as currently implemented, have virtually no mechanisms in place to detect or prevent reputation fraud.

6.1 Summary of Results

In this thesis, we examine two commonly exploited weaknesses of existing reputation systems and discuss various methods of defending against these attacks. The first issue we investigate is the problem of retaliatory negative feedback. Because of the impact negative feedback on a seller's business, many users retaliate by leaving an undeserved negative for the complainer. While reputations have little impact on one's ability to participate as a buyer in the market, buyers do pay attention to their reputations, perhaps because most sellers must spend some time as a buyer in order to pay the market's initiation dues. (15; 95) Leaving feedback (either positive of negative) is a social good with essential zero direct payoff to the rater. Therefore, if sellers are known to retaliate, buyers will hesitate to leave legitimate negative feedback for fear of having their own reputations damaged.

We first examine the effect of retaliation in a simulated evolution experiment. With retaliation allowed, users become unwilling to leave negative feedback preventing the reputation system from discriminating between honest and dishonest users. Ultimately, the market collapses into a non-cooperative state. When retaliation is prohibited or made more difficult through the use of blind feedback, the reputation system is better able to maintain a cooperative state. However, cooperation is only guaranteed when users are forced to leave feedback, suggesting that user apathy toward the reputation system is also a significant problem.

The EM-Trust system of Chapter 4 offers an algorithmic approach to solving the problem of retaliatory negative feedback. Since completely preventing retaliation in real markets is infeasible, EM-Trust operates by assuming that negative feedback may be unreliable. When at least one party to a transaction leaves a negative feedback, all the reputation system knows is that something went

wrong. In such a case, EM-Trust determines the most probable allocation of blame for the failed transaction, rather than simply treating all feedback as completely accurate. In our simulated market, EM-Trust is significantly more capable than eBay's percent positive feedback (PPF) metric at estimating true user honesty in the presence of moderate to high levels of retaliatory negative feedback. It is also able to maintain transaction success rates substantially identical to those of PPF, while reducing the likelihood that a good user will receive a poor reputation and increasing the liquidity of the market.

The second major attack scenario we address is that of the sybil attack. In a sybil attack, a user creates an army of fake user identities and uses them to inflate or reduce a target's reputation. While previous work claims to have solved this problem (68), we demonstrate that this result is due to a poor evaluation methodology. We then introduce Relative Rank, a transformation of the familiar PageRank algorithm (93) that both adapts it for use as a peer-to-peer marketplace reputation system and offers considerably increased resistance to sybil attacks. Finally we demonstrate that a different algorithm in the PageRank family, called RAW, can be made provably sybil proof.

While we cannot address all of the possible ways in which dishonest users may attack, we believe that these results make a significant contribution toward fully robust reputation systems. In particular, these systems both demonstrate that it is possible to make reputation systems more robust against attacks without significant modifications to the feedback collection process. Both systems assume the same simple, positive/negative feedback scheme used by the eBay feedback forum and could be used as drop-in replacements for PPF in existing markets.

6.2 Future Directions

The results of this thesis suggest several avenues for future research, which we look forward to investigating in the future:

Combined systems As currently implemented, EM-Trust and RAW/Relative Rank are two mostly independent systems. In fact, their essential approaches to the problem of robust reputations are quite different: EM-Trust uses a probabilistic model of the feedback process while

RAW/Relative Rank exploits graph structure. One could easily implement both and let the end user form a personal interpretation of the two scores. Ideally, however, we would like to have a joint system that returns reputations that are both sybil-proof and robust against retaliatory negatives. However, the best way to accomplish this combination is not obvious.

Robustness against other attacks In this thesis, we only examine with two common attacks, but there are many other ways in which users game the reputation system. For example, one common strategy is to build a reputation selling low value items, then use it to commit high profit fraud. (62) observed at least two sellers out of 107 that adopted exactly this strategy and then either did not send the purchased goods or sold counterfeit merchandise. Several proposals (64; 84; 117) have included a temporal or memory component to their reputations to help reduce the effectiveness of this attack, but such approaches have not been proven effective. In our opinion, a real solution to this attack will involve not only a method of discounting old feedback, but also a greater understanding of how "portable" a reputation is between items of different types and values. It may also require some method of detecting and discounting aberrant behavior as seen in the temporal dimension much like RAW/Relative Rank discounts sybil-like structures in the trust graph.

Better theories of how and why attacks effect reputation systems While there is a deep literature concerning theoretical models of reputation and their effect on markets, researchers have thus far not investigated the circumstances under which different classes of attacks constitute optimal strategies. In the style of our demonstration in Chapter 3, we would like to see a better theoretical understanding of how these attacks effect the performance of the reputation system in isolation. A better theory of how and why users attack reputation systems will almost surely lead to insights into mechanisms for preventing such abuse.

Standardized evaluations of reputation system performance In many branches of computer science (notably systems and networking), the existence of standard benchmarks and other means of evaluating performance have a very positive effect on progress. The reputation systems community lacks such benchmarks or even a consensus of the best way in which to evaluate competing systems. While we make every attempt to make our simulated market a believable model of real user behavior, we recognize that any such model runs the risk

of oversimplification or bias for or against a particular type of system. A standardized testing framework would provide a neutral venue for comparing systems' relative strengths and weaknesses. Furthermore, such a platform should improve researcher productivity, since it will no longer be necessary to develop a simulator or other benchmarking system along with new reputation systems.

It is our hope that both these suggestions and the new technologies we describe in this thesis eventually find their way into the hands of actual peer-to-peer market users. Peer-to-peer markets offer an enormous selection of goods, many of them difficult to find in traditional retail channels. They also offer opportunities for entrepreneurs to launch small or part-time business that fill important market niches that may not be profitable outside of a peer-to-peer marketplace. In order to energize growth of this valuable sector of the economy, market operators need to work to ensure that their systems truly are as trouble-free and low risk as their marketing claims. More accurate and robust reputations will enable smoother transactions with less risk and will benefit all but the dishonest. We look forward to such a world where robust, effective reputation systems make trading with an individual on the other side of the world as safe, easy, and commonplace as buying from a local shop.

Appendix A

Simulating Online Peer-to-peer Markets

A.1 Introduction

Qualitative testing of new reputation algorithms for large scale peer-to-peer markets presents considerable difficulties, since it is generally not feasible to test experimental reputation systems in real marketplaces. Historical data, if available, can provide an accurate snapshot of user behavior in a real market, but does not allow us to measure the impact of the reputation system on user behavior.

For the research in this thesis, we were unable to obtain such historical data from market operators. "Scraping" a running online market, as done by many of the empirical studies in Chapter 2, is not feasible for our work because of the volume of transactions needed. Most sites actively work to shut down attempts to download such quantities of data.

It is thus necessary to simulate the marketplace if anything more than theoretical results are to be presented, but simulating large peer-to-peer markets is itself a non-trivial problem. The simulation must model the interaction and feedback patterns of real users closely enough that quantitative evaluations of algorithms using the simulator are believable representations of their performance in a real market. At the same time, a simulation is necessarily an abstraction of a real market that can be run tractably on readily available hardware.

In this appendix we discuss the challenges of designing a simulator for peer-to-peer markets. While others have developed marketplace simulations, see for example (117), these simulators are

for smaller markets that do not have many of the observed properties of large scale systems like eBay. We then discuss the problem of choosing parameter values that yield a simulation of eBay with interaction and feedback statistics that closely resemble those reported by (95). Finally we demonstrate the use of simulation in evaluating and comparing reputation system algorithms.

We believe that this simulator achieves our desiderata:

- Its assumptions appear to us to be realistic, at least to a first approximation.
- Its behavior models observed behavior well enough, and robustly enough, so that the results are plausible proxies for what would happen in real marketplaces.
- As far as we can tell, it does not incorporate any biases designed to favor (or disfavor) our models.
- It provides a flexibly large but not overwealming set of tunable parameters, which permits us to explore alternative market models.

The robustness of the simulation results suggests that further complications are unwarranted, at least for our purposes.

Versions of the simulator described in this appendix are used to conduct the experiments presented in Chapters 4 and 5. It is also used as the basis of the evolutionary simulator in the study of retaliatory negative feedback discussed in Chapter 3. However, implementing the evolution mechanism entails considerable changes to the way agent honesty, interactivity, and feedback choices are made. These changes to the simulator framework are described in Chapter 3.

A.2 Overall Design

For simplicity, our market assumes that all agents are trading in a single commodity at a fixed price. Since none of the algorithms we test currently use price, commodity type, or bidding behavior when calculating reputations, it is unnecessary to simulate the full auction process.

The simulator runs for a specified number epochs, each of which consist of a number of individual transactions. Between epochs, the simulator recalculates the reputations of all the agents

in the system. We do not recalculate reputations after each transaction both for efficiency reasons and because feedback is not usually left immediately after completion of a transaction. We find an epoch size of 1000 transactions to be a good tradeoff between the performance cost of evaluating reputations and the negative impact on market behavior of out-of-date reputations. The experiments in this thesis generally run between 50 and 500 epochs.

To simulate a single transaction, the simulator first chooses a seller agent and a buyer agent. The seller and buyer then decide whether they want to interact with each other based on each others' reputations. The simulator continues to choose buyers until it finds a pair that agrees to interact or until a fixed period of time elapses. If a buyer is never found, the seller fails to sell its goods and has to try again later.

Once a buyer/seller pair agrees to interact, the simulator determines their performance in the transaction. Each agent has an honesty parameter (λ_i) indicating their probability of performing acceptably, which is used to randomly generate their performance on each transaction.

Finally, the agents are allowed to leave feedback for each other based on their performance in the transaction. The agents' performance and feedback as well as the transaction time, ID, and other statistics are recorded and the simulator begins the process again for the next transaction.

At the end of an epoch, the simulator runs the reputation system to update agent reputations. It also records some snapshot information about the state of the market to be used for later analysis. Once the specified number of epochs has completed, it records further performance information and exits.

The simulator is controlled by setting the various parameters prior to running a simulation. These parameters and their default values are given in Table A.1.

The simulator is written in Java and allows the user to configure the marketplace parameters, generate sets of agents, and run simulations using either an interactive command line or a batch scripting interface. Most importantly, it is possible to run multiple simulations with different reputation systems on the same initial set of agents.

Simulator function	Parameter name	Default Value
General	Transactions per epoch	1000
	Number of epochs	200
Agent creation	Number of buyers	4000
	Number of sellers	1350
Agent honesty	Proportion of good agents	0.98
	Good agent sub-distribution α	18.0
	Good agent sub-distribution β	2.0
	Bad agent sub-distribution α	2.0
	Bad agent sub-distribution β	18.0
Buyer participation	Mean buy rate	0.2
	Variance of buy rate	0.08
	Mean sell rate	008
	Variance of sell rate	008
Seller participation	Mean buy rate	0.08
	Variance of buy rate	0.0128
	Mean sell rate	64
	Variance of sell rate	1.024
Interaction	Interaction threshold	0.884
	Threshold width	0.2
Respawning	New agent creation rate	25.0
	Respawn rate	0.6
Feedback	First feedback probability for good agents	0.3
	Second feedback probability for good agents	0.6
	First feedback probability for bad agents	0.1
	Second feedback probability for bad agents	0.5
Retaliation	Good agent retaliation rate	0.25
	Bad agent retaliation rate	0.75
Bayesian prior	γ (Mixture ratio)	0.98
	α_1	18.0
	β_1	2.0
	α_2	2.0
	β_2	18.0

Table A.1: Description of simulator parameters and their default values.

A.3 Agents

Before the simulator can begin to process transactions, it must create the pool of agents that will participate in the market. Each agent has a set of characteristics, some of which are chosen by the user while others are sampled random from user-parameterized distributions.

A.3.1 Agent Characteristics

Agents in our marketplace have a disposition, good or bad, and a type, buyer or seller. Good agents are mostly honest and have low failure rates. Bad agents are assumed to be dishonest or incompetent and have high failure rates. The type determines whether the agent primarily buys or sells items in the market. The initial number of buyer and seller agents is specified by the user.

An agent's type and disposition influence its characteristics:

Honesty The honesty characteristics indicates the agent's probability of successfully completing its half of a transaction. The term "honesty" is a misnomer — this value governs the rate of failures for all possible reasons, not simply dishonesty.

Buying and selling rates These characteristics govern the Poisson processes associated with each agent that generate buy and sell offer times. Sellers typically have higher sell rates, while buyers have higher buy rates. However, both rates can be non-zero, so some agents will both buy and sell items in the market. Originally, interaction rates were sampled from an exponential distribution, but in later versions of the simulator, we use the more realistic continuous Mandelbrot distribution.

Feedback and retaliation rates Each agent has three characteristics controlling how it leaves feedback: its probability of leaving feedback before it receives feedback from its partner, its probability of leaving a feedback after its partner, and its probability of retaliating for negative feedback.

Interaction threshold and width Given a potential partner's reputation, these characteristics control how likely an agent is to interact with the partner.

The values of these characteristics are determined at the time the agent is created. Each agent has a unique honesty, buy rate, and sell rate value that are randomly generated by distributions controlled by simulation parameters and agent type and disposition. The feedback and interaction characteristics' values are the same for all agents of a particular type and disposition, and are simply copied from global simulator parameters.

A.3.2 Creating Agents

The simulator creates new agents at three points in the simulation: at startup, periodically during simulation to model new people joining the market, and when agents "respawn" or discard their identities because of a low reputation. Agents only leave the market during the respawning process.

The user chooses the size of the initial pool of agents by specifying the number of each type to create. These agents' dispositions are sampled from the honesty distribution and have interactivity and feedback characteristics as specified by their type and disposition.

While the simulation is running, new agents occasionally join the market. The rate of new agent creation is controlled by the new agent creation rate parameters, which specifies the percentage growth in the market per simulated day. The simulator chooses a random type for new agents by sampling from the distribution determined by the mix of types in the initial pool. The new agents' other characteristics are determined in the same fashion as the initially created ones.

A.3.3 Agent Respawning

In real marketplaces, agents whose reputations fall too low will likely discard their identity and re-enter the system as a new user. This non-persistence of identities led Zacharia et al (118) to suggest that reputation systems should never allow reputation to fall below the level of a new user. While eBay's Percent Positive Feedback and the systems discussed in this thesis follow this guideline, the fact remains that a new user with a zero reputation may receive more trust than a user that has zero reputation after a number of transactions.

At any point in the simulation, if an agent finds that it has a higher probability of interacting as

a new agent than it does with its current reputation, it will discard its old identity and either leave the market or rejoin as a new agent, a process we call "respawning." The respawn rate parameter controls the rate at which agents respawn versus leave under such circumstances.

We simulate creating a new identity by adding a new agent with the same parameter values to the simulated marketplace. The old identity is marked inactive and is removed from the buying and selling queues. The simulator tracks these identity changes so that we can analyze the frequency with which individual agents change identities.

A.4 Generating Transactions

We model an agent's participation in the market with a pair of Poisson processes, one for buying and one for selling. The rate of these processes is controlled by the agent's buying and selling rate parameters. These processes generate a series of times when an agent wants to buy or sell in the market.

The simulator maintains two priority queues each containing all active agents sorted by their next buy or sell time. When conducting a transaction, the simulator chooses a seller from the front of the seller queue, then pulls potential buyers off the buyer queue until an agreeable match is found. An agent that does not find a partner within four simulated days of initially offering to buy or sell will give up, causing the offer to expire. After either a completed transaction or when a buy or sell offer expires, the simulator generates a new buy or sell time as appropriate and inserts the agents back in the priority queues.

A.4.1 Transaction Rate Distributions

Agents in real markets exhibit a wide range of interaction frequencies. Typically, these frequencies are determined by factors external to the market. Resnick and Zeckhauser (95) found that the median seller sold 50 items in the 100 day period of their study and the median buyer bought 20 items in the same period. We use these values to determine the distribution of buyer and seller rates in our simulator.

But which distribution to use? Looking at the interactivity histograms in (95), it is clear that the number of agents that buy or sell at a given rate tends to decrease as the rate increases. The tails of the distribution are also quite heavy.

Earlier versions of our simulator use an exponential distribution, with the well-known density function:

$$f(x;k) = ke^{-kx}$$

We can easily find values of k that give the appropriate medians: 0.0139 for sellers and 0.0347 for buyers. The exponential distribution seems reasonable for small rates; however, its tails are not fat enough. The 99th percentile point of the exponential for sellers is approximate 332 items per 100 day period (3.32 items/day) and the 99.9% point is 497 (4.97 items/day). (95) found considerably more than 1% of sellers sold at least 403 items in a 100 day period. Despite its deviation from reality, we use the exponential distribution in the version of the simulator used for the experiments in Chapter 4. In that version of the simulator, four parameters control the distribution of agent interactivity: one each for the buyers' buy rate, buyers' sell rate, sellers' buy rate, and sellers' sell rate.

The problem with the exponential distribution, of course, is that the probability density drops off exponentially fast. A better alternative is one of the power-law distributions, like the Zipf or Zeta, long tailed distributions which tend to be a good fit to many Internet phenomena. Since the Zipf and Zeta distributions are discrete, we need to generalize them to handle continuous interaction rates. A continuous analog to the Zeta distribution can be straightforwardly shown to have density function:

$$f(x;k,x_m) = \frac{kx_m^k}{x^{k+1}}$$

with parameters x_m, the minimum value that x can take, and k, which controls the rate at which the density decays with increasing x. For our purposes, we fix $x_m = 1$. Agents that do not interact even once during a 100 day period are of little interest because the simulations typically require less than 100 simulated days to complete. The median of this continuous Zeta is:

$$P_{0.5} = x_m 2^{1/k}$$

so it is easy to estimate k given the known medians. Unfortunately, this distribution has the opposite

problem of the exponential: its tails are too fat. If our median seller sells an item every two days, then the top 1% of sellers will sell nearly 2 billion items per day. Clearly, this rate is ridiculously unrealistic.

Our solution is to use a continuous version of the Mandelbrot distribution, a generalization of the Zipf/Zeta distributions (82). This distribution allows us to have a rather shallow decrease in density up to a certain point, then a much steeper decrease as the rate goes towards infinity. Its density function is:

$$f(x; x_m, r, k) = \frac{k(x_m + r)^k}{(x + r)^{k+1}}$$

Because we have two parameters to estimate (once again, we set $x_m = 1$), we use both the median and the 99th percentile point to determine the parameter values. For the 99% points, we use $P_{0.99} = 1000$ for sellers and $P_{0.99} = 400$ for buyers. These admittedly arbitrary choices correspond to rates of 10 sales/day and 4 buys/day, which impressionistically seem to be reasonable rates for highly interactive agents. More detailed information about interaction rates in real markets would permit us to make more accurate estimates.

Given these points, we obtain the following estimates for the distribution parameters:

	x_m	k	r
Buyers' buy rate	1	2.0510	46.2539
Sellers' sell rate	1	2.0977	124.1356

We would require eight parameters to describe all four interaction rates; however, in the experiments of Chapters 3 and 5, we fix buyers' sell rates and sellers' buy rates to zero. Earlier experiments show that the small number of transactions due to these rates have negligible effect on reputation system performance, so we ignore them for simplicity.

The differences between these three approaches to modeling agent interaction can be seen in Figure A.1. This graph plots the probability densities of the three distributions with parameters chosen to model sellers as observed in (95).

To generate agent interaction rates, we need a way of generating random numbers with this continuous Mandelbrot distribution. It can be easily shown that if

$$X = \frac{x_m + r}{(1 - U)^{1/k}} - r$$

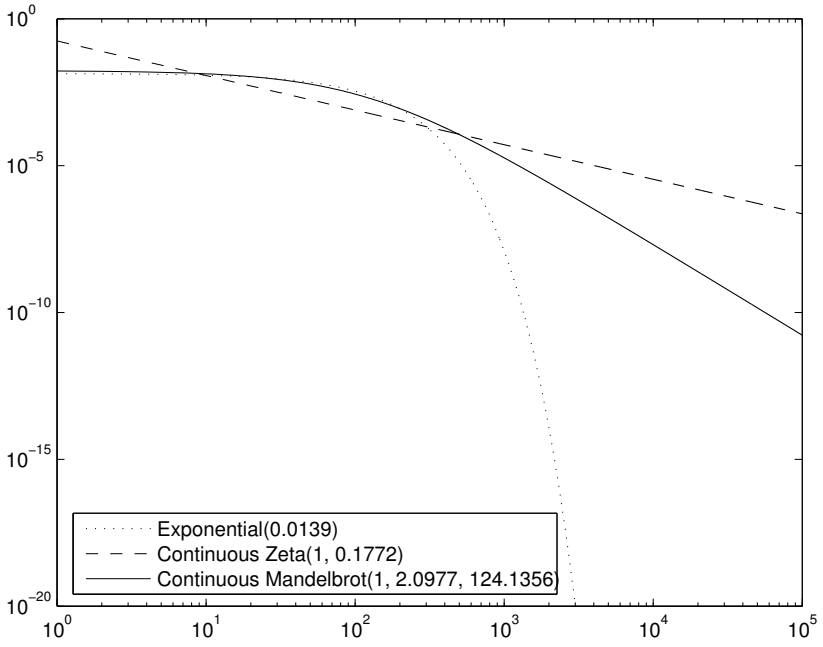


Figure A.1: Comparison of Exponential, Continuous Zeta, and Continuous Mandelbrot distributions all with median 50.

where U is a Uniform(0,1) distributed random variable, then X is continuous Mandelbrot(x_m, k, r) distributed. Since we choose parameters in terms of 100 day periods, we simply divide X by 100 to obtain an interaction rate in terms of items/day.

A.5 Agent Interactivity

Once a potential buyer/seller pair are generated using the buy/sell Poisson processes, the agents are given the choice of whether they want to interact with each other.

Two global parameters determine how agents choose whether to interact or not. The first parameter, the interaction threshold, determines the reputation value of its partner above which the agent is likely to want to interact. The second parameter, the transition width, is the width of the transition region between always wanting to interact and always declining to interact. The probability of

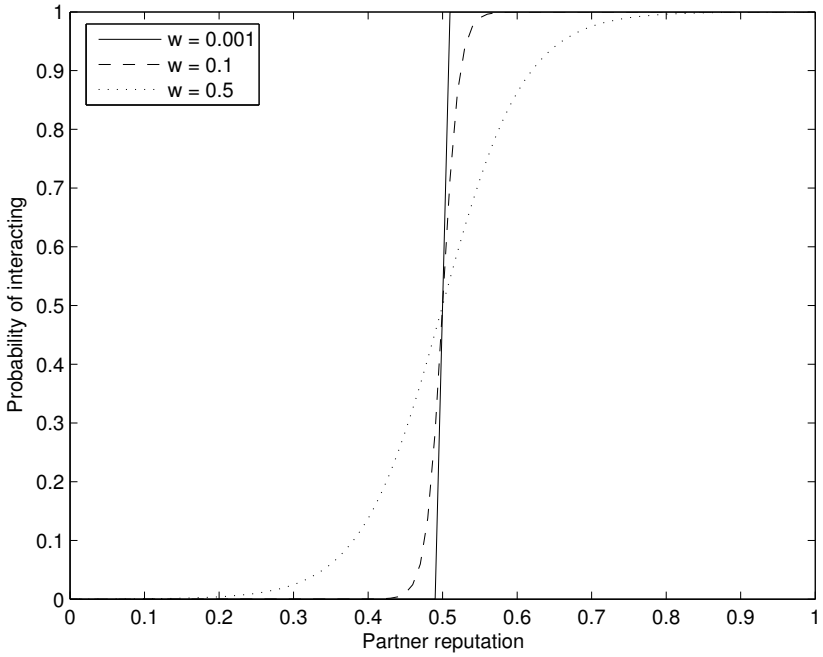


Figure A.2: The interactivity function with threshold 0.5 and three different interaction widths.

interaction is given by a form of the logistic function:

$$I(r) = \frac{1}{1 + e^{[-\frac{2 \ln 99}{w}(r-t)]}}$$

where r is the potential partner's reputation, t is the interaction threshold and w is the transition width.

In this function, the interaction threshold controls the value of r for which the probability of interaction is 0.5. The transition width controls the steepness of the slope of the logistic function and is defined as the length of the interval of r values where $I(r)$ lies between 0.01 and 0.99. Figure A.2 plots examples of this function at different interaction widths.

To determine whether to interact with a potential partner, an agent feeds the partner's reputation to this function to obtain a probability of interaction. It then generates a Bernoulli distributed random variable with this probability to determine whether to interact or not. If the transition width is set to 0, the interaction is no longer stochastic and the agent always interacts if the partner's reputation is greater than the interaction threshold and declines to interact otherwise.

Agents start with zero reputations in the eBay and standard EM-Trust systems — Bayesian EM-Trust starts users at the mean of the reputation prior — so agents will never start to interact unless we handle interaction with new users specially. Therefore, an additional parameter, the probability of interacting with new users controls how often agents choose to interact with others that do not yet have a reputation.

In the version of the simulator used for Chapters 4 and 5, the simulator allows the user to specify the interaction threshold, transition width, and new user interactivity globally but does not permit individual agents to have their own parameter values. The evolutionary simulator of Chapter 3 extends this basic framework considerably: each agent has probability of interacting with new users and two interaction thresholds, one when dealing with partners who have received less than 10 feedbacks and a separate one for partners with more experience.

A.6 Leaving Feedback

After the two partners decide to interact with each other, the simulator randomly generates their performance for the transaction by sampling Bernoulli distributions parameterized by the agents' honesties. The agents may then leave feedback for each other. The feedback each agent leaves is influenced by both agents' performance and three parameters: the probability of leaving the first feedback, the probability of leaving a second feedback, and the probability of leaving a retaliatory negative. Agents have different feedback strategies depending on their disposition but all agents of a given disposition share common parameters.

After the agent's behavior in a transaction has been recorded, the simulator polls each agent to see if they wish to leave the first feedback. If neither agent wants to leave the first feedback, neither leave feedback for the transaction. If both agree to leave the first feedback, the simulator picks one or the other randomly with equal probability.

The agent that leaves feedback first must determine the type of feedback to leave solely on the basis of its own behavior and its partner's behavior. Since the other agent has not yet left a feedback, it cannot base its decision on the other's feedback.

Good agents always leave accurate first feedback: if the other agent behaved correctly, it will leave a positive, otherwise it will leave a negative. A bad agent will not leave a positive feedback if it behaved dishonestly in the transaction because leaving a positive will allow its partner to leave a negative without fear of retribution. Instead, it may leave a pre-emptive negative feedback to try to disguise its responsibility for a bad transaction. If it behaved honestly, it will leave an accurate feedback.

While it is a tunable parameter, all of our tests are run with bad agents having a low probability of leaving the first feedback. We believe the optimal strategy for someone running a scam in a peer-to-peer market is to leave no feedback unless it receives a negative, in which case it retaliates. Using this strategy, in the best case, no feedback is left, so there is no evidence of the scam.

After the first agent leaves feedback, the simulator queries the other agent to see if it wants to leave the second feedback. If the agent wants to leave a second feedback, it can base its decision on the behavior of the agents as well as the feedback left by the first agent. If either a good or bad agent receives a positive feedback, and it decides to leave any feedback, it will leave an accurate second feedback.

If a good agent receives a negative feedback it will leave a retaliatory negative feedback according to its probability of retaliation, regardless of its own behavior or that of its partner. If it does not choose to retaliate, the good agent will leave an accurate feedback. Except for retaliation, good agents do not generally try to game the feedback system because we assume their failures are due more to mistakes or lack of competence than to malice.

Bad agents will also retaliate for negative feedback according to their probability of retaliation. If they do not retaliate, they will always leave a negative feedback if the other agent did not behave correctly. However, a bad agent will leave a positive feedback for a correctly behaving partner only if the agent itself also behaved correctly. Otherwise, it will leave no feedback. We assume bad agent's behave badly chiefly out of dishonesty so they try to use the feedback system to cover their tracks as much as possible.

A.7 Discussion

For various reasons, it is usually not feasible to test experimental reputation system algorithms in real peer-to-peer markets. However, theoretical results and small worked examples can only go so far toward showing the strengths and weaknesses of new approaches to trust management. Simulation allows researchers to make quantitative evaluations and comparisons of reputation systems without access to a real large-scale peer-to-peer market.

Our simulator design achieves a balance between the competing goals of simplicity and realism. While the total number of parameters is manageable, with proper tuning we can achieve interaction and feedback distributions that closely resemble observations of real marketplaces. In addition to simulating real markets, parameters can be tweaked to explore alternate, hypothetical scenarios. Running multiple simulations in differently configured markets makes it easier to develop reputation systems that are robust to the sort of changing conditions that occur in real marketplaces.

While certainly not a replacement for testing in real markets, simulation can greatly assist the creation of new reputation systems. By allowing researchers to explore the performance of their algorithms quickly and easily, the development cycle is shortened and more optimization is possible. It is our hope that increasingly realistic simulation will enable the creation of the next generation of reputation systems for real peer-to-peer markets.

Bibliography

[1] E. Adar and B. Huberman, "Free riding on Gnutella," *First Monday*, vol. 5, no. 10, October 2000, http://www.firstmonday.org/issues/issue5_10/adar/index.html.

[2] J. Ahuja, "The age of eBay and e-commerce," *CELCEE Digest*, no. 04-08, December 2004, http://www.celcee.edu/publications/digest/Dig04-08.html?version=print.

[3] G. A. Akerlof, "The market for "lemons": Quality, uncertainty and the market mechanism," *Quarterly Journal of Economics*, vol. 84, no. 3, pp. 488–500, August 1970.

[4] S. Alanet, "Trust metrics (idea)," September 2002, http://everything2.com/index.pl?node_id=9229.

[5] All Enthusiast, Inc., "ResellerRatings.com," http://www.resellerratings.com/.

[6] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *Science*, vol. 211, no. 4489, pp. 1390–1396, 1981.

[7] P. Bajari and A. Hortaçsu, "The winner's curse, reserve prices, and endogenous entry: Empirical insight from eBay auctions," *RAND Journal of Economics*, vol. 34, no. 2, pp. 329–355, Summer 2003.

[8] K. S. Barber, K. Fullam, and J. Kim, "Challenges for trust, fraud and deception research in multi-agent systems," *Trust, Reputation, and Security: Theories and Practice*, pp. 8–14, 2003.

[9] R. Bhattacharjee and A. Goel, "Avoiding ballot stuffing in eBay-like reputation systems," in *Proceedings of the 3rd Workshop on Economics of Peer-to-Peer Systems (P2PECON)*, 2005.

[10] M. Bianchini, M. Gori, and F. Scarselli, "Inside PageRank," *ACM Transactions on Internet Technology*, vol. 5, no. 1, February 2005.

[11] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the of World Wide Web '98 Conference*, 1998.

[12] D. Bryan, D. Lucking-Reiley, N. Prasad, and D. Reeves, "Pennies from eBay: the determinants of price in online auctions," Econometric Society, Econometric Society World Congress 2000 Contributed Papers 1736, Aug. 2000, available at http://ideas.repec.org/p/ecm/wc2000/1736.html.

[13] S. Buchegger and J.-Y. L. Boudec, "A robust reputation system for p2p and mobile ad-hoc networks," in *Proceedings of the 2nd Workshop on Economics of Peer-to-Peer Systems (P2PECON)*, 2004.

[14] ——, "Self-policing mobile ad-hoc networks by reputation systems," *IEEE Communications Magazine*, July 2005.

[15] L. Cabral and A. Hortaçsu, "The dynamics of seller reputation: Theory and evidence from eBay," *Discussion Paper Series: Centre for Economic Policy Research*, no. 4345, 2004.

[16] M. M. Calkins, "My reputation always had more fun than me: The failure of eBay's feedback model to effectively prevent online auction fraud," *The Richmond Journal of Law and Technology*, vol. 7, no. 4, Spring 2001, http://law.richmond.edu/jolt/v7i4/note1.html.

[17] M. Capecchi, "Altering the genome by homologous recombination," *Science*, vol. 244, pp. 1288–1292, 1989.

[18] I. Chakraborty and G. Kosmopoulou, "Auctions with shill bidding," *Economic Theory*, vol. 24, pp. 271–287, 2004.

[19] A. Chavez and P. Maes, "Kasbah: An agent marketplace for buying and selling goods," in *Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM '96)*, 1996.

[20] T. J. Chemmanur and P. Fulghieri, "Reputation, renegotiation, and the choice between bank loans and publicly traded debt," *The Review of Financial Studies*, vol. 7, no. 3, pp. 475–506, Fall 1994.

[21] A. Cheng and E. Friedman, "Sybilproof reputation mechanisms," in *Proceedings of the SIGCOMM '05 P2P-ECON Workshop*, August 2005.

[22] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," August 2003, Yale SOM Working Paper No's. ES-28 & MK-15. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=432481.

[23] A. Clausen, "The cost of attack of PageRank," in *Proceedings of the International Conference on Agents, Web Technologies and Internet Commerce (IAWTIC)*, 2004.

[24] B. Cohen, "Incentives build robustness in bit torrent," in *Proceedings of the 1st Workshop on Economics of Peer-to-peer Systems (P2PECON)*, 2003.

[25] W. Dai and H. Finney, "Towards a theory of reputation," from the Cypherpunks archive available at http://cypherpunks.venona.com/date/1995/11/msg01043.html.

[26] C. Dellarocas, "Building trust online: The design of robust reputation reporting mechanisms," in *Information Society or Information Economy? A Combined Perspective on the Digital Era*, G. Doukidis, N. Mylonopoulos, and N. Pouloudi, Eds. Idea Book Publishing, 2003, ch. 7, pp. 95–113.

[27] ——, "The digitization of word of mouth: Promise and challenges of online feedback mechanisms," *Management Science*, vol. 49, no. 10, pp. 1407–1424, 2003.

[28] S. Dewan and V. Hsu, "Adverse selection in electronic markets: Evidence from online stamp auctions," *Journal of Industrial Economics*, vol. 52, no. 4, pp. 497–516, December 2004.

[29] D. W. Diamond, "Reputation acquisition in debt markets," *Journal of Political Economy*, vol. 97, no. 4, pp. 828–862, 1989.

[30] J. R. Douceur, "The sybil attack," in *Proceedings of the IPTPS02 Workshop*, 2002.

[31] eBay, Inc., "Detailed seller ratings," http://pages.ebay.com/help/feedback/detailed-seller-ratings.html.

[32] ——, "eBay.com," http://www.ebay.com.

[33] ——, "Feedback revision process," http://pages.ebay.com/community/suggestion/feedback-results.html.

[34] ——, "Resolving feedback disputes," http://pages.ebay.com/help/feedback/feedback-disputes.html.

[35] ——, "Shopping.com," http://www.shopping.com/.

[36] ——, "Form 10-k: 2005 annual report," 2005, http://edgar.sec.gov/Archives/edgar/data/1065088/000095013406003678/f17187e10vk.htm.

[37] G. Ellison and S. F. Ellison, "Lessons about markets from the Internet," *Journal of Economic Perspectives*, vol. 19, no. 2, pp. 139–158, 2005.

[38] ——, "Search, obfuscation, and price elasticities on the Internet," February 2005, working paper. http://econ-www.mit.edu/faculty/download_pdf.php?id=942.

[39] Escrow.com, "Escrow.com," https://www.escrow.com/index.asp.

[40] Federated Department Stores, Inc., "2005 annual report," 2005, http://www.fds.com/ir/ann.asp.

[41] M. Feldman, K. Lai, I. Stoica, and J. Chuang, "Robust incentive techniques for peer-to-peer networks," in *Proceedings of the ACM E-Commerce Conference (EC '04)*, May 2004.

[42] M. Feldman and J. Chuang, "The evolution of cooperation under cheap pseudonyms," in *Proceedings of the 7th International IEEE Conference on E-Commerce Technology (CEC '05)*, July 2005.

[43] A. Fernandes, E. Kotsovinos, S. Östring, and B. Dragovic, "Pinocchio: Incentives for honest participation in distributed trust management," in *Proceedings of the 2nd International Conference on Trust Management (iTrust 2004)*, March 2004.

[44] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós, "Toward scaling fully personalized PageRank: Algorithms, lower bounds, and experiments," *Internet Mathematics*, vol. 2, no. 3, pp. 333–358, 2005.

[45] E. Friedman and P. Resnick, "The social cost of cheap pseudonyms," *Journal of Economics and Management Strategy*, vol. 10, pp. 173–199, 2001.

[46] Friendster, Inc., "Friendster," http://www.friendster.com.

[47] M. Froomkin, "The internet as a source of regulatory arbitrage," in *Borders in Cyberspace*, B. Kahin and C. Nesson, Eds. MIT Press, 1997.

[48] E. Garfield, "Citation analysis as a tool in journal evaluation," *Science*, no. 178, pp. 471–479, 1972.

[49] A. Ghose, P. G. Ipeirotis, and A. Sundararajan, "Reputation premiums in electronic peer-to-peer markets: Analyzing textual feedback and network structure," in *Proceedings of the SIGCOMM '05 P2P-ECON Workshop*, 2005.

[50] R. Gibbons, *Game Theory for Applied Economists*. Princeton, NJ: Princeton University Press, 1992.

[51] C. Goffman, "And what is your Erdös number?" *American Mathematical Monthly*, vol. 76, 1969.

[52] J. Golbeck and J. Hendler, "Accuracy of metrics for inferring trust and reputation in semantic web-based social networks," in *Proceedings of EKAW*, 2004.

[53] ——, "Reputation network analysis for email filtering," in *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, July 2004.

[54] A. V. Goldberg, "Efficient graph algorithms for sequential and parallel computers," Ph.D. dissertation, MIT, 1987.

[55] A. V. Goldberg and R. E. Tarjan, "A new approach to the maximum-flow problem," *Journal of the ACM*, vol. 35, no. 4, pp. 921–940, 1988.

[56] J. L. Goldstein, "Laskers for 2001: Knockout mice and test-tube babies," *Nature Medicine*, vol. 7, no. 10, pp. 1079–1080, 2001.

[57] Q. He, D. Wu, and P. Khosla, "Sori: A secure and objective reputation-based incentive scheme for ad-hoc networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC2004)*, 2004.

[58] T. Hogg and L. Adamic, "Enhancing reputation mechanisms via online social networks," in *Proceedings of the 5th ACM Conference on Electronic Commerce*, 2004.

[59] T. Hossain and J. Morgan, "...Plus shipping and handling: Revenue (non) equivalence in field experiments," *The B.E. Journal of Economic Analysis and Policy*, vol. 6, no. 2, January 2006.

[60] D. Houser and J. Wooders, "Reputation in auctions: Theory, and evidence from eBay," *Journal of Economics and Management Strategy*, vol. 15, no. 2, pp. 353–369, Summer 2006.

[61] D. Hughes, G. Coulson, and J. Walkerdine, "Free riding on Gnutella revisited: The bell tolls?" *IEEE Distributed Systems Online*, 2005.

[62] G. Z. Jin and A. Kato, "Price, quality and reputation: Evidence from an online field experiment," December 2005, working paper. To appear in the RAND Journal of Economics. http://www.glue.umd.edu/ ginger/research/ebay-exp-dec0105.pdf.

[63] K. John and D. C. Nachman, "Risky debt, investment incentives, and reputation in a sequential equilibrium," *The Journal of Finance*, vol. 40, no. 3, pp. 863–878, July 1985.

[64] A. Josang and R. Ismail, "The Beta reputation system," in *Proceedings of the 15th Bled Conference on Electronic Commerce*, June 2002.

[65] J. Joseph E. Harrington, "The role of party reputation in the formation of policy," *Journal of Public Economics*, vol. 49, pp. 107–121, 1992.

[66] J. M. R. Jr., "Trust in electronic markets: The convergence of cryptographers and economists," *First Monday*, vol. 1, no. 2, August 1996, http://www.firstmonday.org/issues/issue2/index.html.

[67] S. Jun and M. Ahamad, "Incentives in BitTorrent induce free riding," in *Proceedings of the 3rd Workshop on Economics of Peer-to-peer Systems (P2PECON)*, 2005.

[68] S. Kamvar, M. Scholsser, and H. Garcia-Molina, "The EigenTrust algorithm for reputation management in p2p networks," in *Proceedings of the 12th International World Wide Web Conference*, 2003.

[69] J. Kennes and A. Schiff, "The value of a reputation system," Economics Working Paper Archive at WUSTL, Tech. Rep. 0301011, Jan. 2003, http://ideas.repec.org/p/wpa/wuwpio/0301011.html.

[70] R. Khare and A. Rifkin, "Weaving a web of trust," *World Wide Web Journal*, vol. 2, no. 3, pp. 77–112, 1997.

[71] T. Khopkar, X. Li, and P. Resnick, "Self-selection, slipping, salvaging, slacking, and stoning: the impacts of negative feedback at eBay," in *Proceedings of the 2005 ACM Electronic Commerce Conference (EC '05)*, 2005.

[72] G. Kosmopoulou and D. G. D. Silva, "The effect of shill bidding upon prices: Experimental evidence," EconWPA, Tech. Rep. 0512002, Dec. 2005.

[73] E. Kotsovinos, P. Zerfos, and N. M. Piratla, "Jiminy: A scalable incentive-based architecture for improving rating quality," in *Proceedings of the 4th International Conference on Trust Management (iTrust)*, 2006.

[74] D. M. Kreps, P. Milgrom, J. Roberts, and R. Wilson, "Rational cooperation in the finitely repeated prisoners' dilemma," *Journal of Economic Theory*, vol. 27, pp. 245–252, 1982.

[75] D. M. Kreps and R. Wilson, "Reputation and imperfect information," *Journal of Economic Theory*, vol. 27, pp. 253–279, 1982.

[76] ——, "Sequential equilibrium," *Econometrica*, vol. 50, no. 4, pp. 863–894, July 1982.

[77] J. Legon, "'Phishing' scams reel in your identity," *CNN.com*, January 2004.

[78] G. Lewis, "Asymmetric information, adverse selection and seller revelation on eBay motors," Ph.D. dissertation, University of Michigan, 2006.

[79] LinkedIn Corporation, "LinkedIn," http://www.linkedin.com.

[80] D. Lucking-Reiley, D. Bryan, N. Prasad, and D. Reeves, "Pennies from eBay: The determinant in online auctions," 2006, manunscript. To appear in *Journal of Industrial Economics*. Cited version available at http://www.u.arizona.edu/ dreiley/papers/PenniesFromEBay.html.

[81] G. J. Mailath, "Who wants a good reputation?" *Review of Economic Studies*, vol. 68, pp. 415–441, 2001.

[82] B. B. Mandelbrot, *The Fractal Geometry of Nature*. New York: W.H. Freeman, 1983.

[83] M. Marino, "eBay's revised feedback policy," *AuctionBytes-Update*, no. 5, January 2000, http://www.auctionbytes.com/cab/abu/y200/m01/abu0005/s06.

[84] S. Marsh, "Formalising trust as a computational concept," Ph.D. dissertation, University of Stirling, 1994.

[85] S. Marti, T. Giuli, K. Lai, and M. Baker, "Mitigating routing misbehavior in mobile ad hoc networks," in *Proceedings of MOBICOMM 2000*, 2000, pp. 255–265.

[86] P. Michiardi and R. Molva, "Core: A collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks," in *Proceedings of the 6th IRP Conference on Secure Communication and Multimedia*, 2002.

[87] P. Milgrom, "Auctions and bidding: A primer," *Journal of Economic Perspectives*, vol. 3, no. 3, pp. 3–22, Summer 1989.

[88] P. Milgrom and J. Roberts, "Predation, reputation, and entry deterrence," Center for Research on Organizational Efficiency, Stanford University, Tech. Rep. 353, 1981.

[89] H. Moed, "Citation analysis of scientific journals and journal impact measures," *Current Science*, vol. 89, no. 12, December 1990.

[90] MySpace.com, "MySpace.com,," http://www.myspace.com.

[91] M. A. Nowak and K. Sigmund, "Evolution and indirect reciprocity by image scoring," *Nature*, vol. 393, 1998.

[92] A. Ockenfels, D. H. Reiley, and A. Sadrieh, "Online auctions," 2006, working paper to appear in *Handbook of Information Systems and Economics*.

[93] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the web," Stanford University, Tech. Rep. 1999-66, 1999.

[94] G. Pinski and F. Narin, "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics," *Information Processing and Management*, vol. 12, pp. 297–312, 1976.

[95] P. Resnick and R. Zeckhauser, "Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system," in *The Economics of the Internet and E-Commerce*, M. R. Baye, Ed. Elsevier Science, 2002, volume 11 of Advances in Applied Microeconomics.

[96] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara, "Reputation systems: Facilitating trust in Internet interactions," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, December 2000.

[97] P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood, "The value of reputation on eBay: A controlled experiment," January 2006, working paper. To appear in Experimental Economics. http://www.si.umich.edu/ presnick/papers/postcards/.

[98] R. Richmond, "Online firms see small businesses as the next big Internet market," *The Wall Street Journal*, November 2002.

[99] R. Selten, "The chain store paradox," *Theory and Decision*, vol. 9, no. 2, pp. 127–159, April 1978.

[100] Slashdot.org, "Slashdot FAQ: Comments and moderation," http://slashdot.org/faq/com-mod.shtml.

[101] I. Steiner, "AuctionBytes survey reveals eBay phishing, account hijacking problems," *AuctionBytes-Update*, September 2006.

[102] ——, "eBay's 'Feedback 2.0' revealed," *AuctionBytes-Newsflash*, no. 1458, January 2007.

[103] S. Tadelis, "What's in a name? Reputation as a tradeable asset," *The American Economic Review*, vol. 89, no. 3, pp. 548–563, June 1999.

[104] ——, "The market for reputations as an incentive mechanism," *Journal of Political Economy*, vol. 110, no. 4, pp. 854–882, 2002.

[105] J. Traupman, "EM-Trust: A robust reputation algorithm for peer-to-peer marketplaces," University of California, Berkeley, Computer Science Division, Tech. Rep. UCB/CSD-05-1400, July 2005.

[106] ——, "Resisting sybils in peer-to-peer markets," 2007, to appear in the *Proceedings of the Joint iTrust and PST Conference on Privacy, Security, and Trust Management*.

[107] J. Traupman and R. Wilensky, "Robust reputations for peer-to-peer marketplaces," in *Proceedings of the 4th International Conference on Trust Management (iTrust)*, 2006.

[108] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," *The Journal of Finance*, vol. 16, no. 1, pp. 8–37, March 1961.

[109] L. von Ahn, M. Blum, and J. Langford, "Telling humans and computers apart (automatically)," Carnegie Mellon University, School of Computer Science, Tech. Rep. CMU-CS-02-117, 2002.

[110] L. Walker and A. Klein, "eBay opens web site to retailers," *The Washington Post*, June 2001.

[111] K. Walsh and E. G. Sirer, "Fighting peer-to-peer spam and decoys with object reputation," in *Proceedings of the 3rd Workshop on Economics of Peer-to-peer Systems (P2PECON)*, 2005.

[112] W. Wang, Z. Hidvégi, and A. B. Whinston, "Shill bidding in multi-round online auctions," in *Proceedings of the 35th Hawaii International Conferences on Systems Sciences*, 2002.

[113] S. Wasserman, K. Faust, and D. Iaccobucci, *Social Network Analysis: Methods and Applications*. Cambrige University Press, 1995.

[114] D. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton, NJ: Princeton University Press, 2003.

[115] J. Wilkinson, "Understanding eBay's new 'mutual feedback withdrawal' policy," *AuctionBytes-Update*, no. 117, April 2004, http://www.auctionbytes.com/cab/abu/y204/m04/abu0117/s04.

[116] N. Wingfield, "Top eBay vendors unite to voice their concerns," *Startup Journal*, May 2004, http://startup.wsj.com/ecommerce/ecommerce/20040526-wingfield.html.

[117] B. Yu and M. Singh, "A social mechanism of reputation management in electronic communities," in *Proceedings of the 4th International Workshop on Cooperative Information Agents*, 2000.

[118] G. Zacharia, A. Moukas, and P. Maes, "Collaborative reputation mechanisms in electronic markets," in *Proceedings of the 32nd Hawaii International Conference on System Sciences*, 1999.

[119] M. Zuckerberg, "Facebook," http://www.facebook.com.

All web content retrieved and archived on May 16, 2007.