Computational Models of Trust and Reputation: Agents, Evolutionary Games, and Social Networks

by

Lik Mui

B.S., M.Eng., Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 1995

M. Phil., Management Studies, Department of Social Studies
University of Oxford, 1997

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

December 20, 2002

Signature of Author: ____
Department of Electrical Engineering and Computer Science
December 20, 2002

Certified by: ____
Peter Szolovits
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: ____
Arthur C. Smith
Professor of Electrical Engineering and Computer Science
Chairman, Department Committee on Graduate Students

Computational Models of Trust and Reputation: Agents, Evolutionary Games, and Social Networks

by

Lik Mui

Submitted to the Department of Electrical Engineering and Computer Science
on December 20, 2002 in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy in
Electrical Engineering and Computer Science

Abstract

Many recent studies of trust and reputation are made in the context of commercial reputation or rating systems for online communities. Most of these systems have been constructed without a formal rating model or much regard for our sociological understanding of these concepts.

We first provide a critical overview of the state of research on trust and reputation. We then propose a formal quantitative model for the rating process. Based on this model, we formulate two personalized rating schemes and demonstrate their effectiveness at inferring trust experimentally using a simulated dataset and a real world movie-rating dataset. Our experiments show that the popular global rating scheme widely used in commercial electronic communities is inferior to our personalized rating schemes when sufficient ratings among members are available. The level of sufficiency is then discussed. In comparison with other models of reputation, we *quantitatively* show that our framework provides significantly better estimations of reputation. "Better" is discussed with respect to a rating process and specific games as defined in this work.

Secondly, we propose a mathematical framework for modeling trust and reputation that is rooted in findings from the social sciences. In particular, our framework makes explicit the importance of *social information* (*i.e.*, indirect channels of inference) in aiding members of a social network choose whom they want to partner with or to avoid. Rating systems that make use of such indirect channels of inference are necessarily personalized in nature, catering to the individual context of the rater.

Finally, we have extended our trust and reputation framework toward addressing a fundamental problem for social science and biology: evolution of cooperation. We show that by providing an indirect inference mechanism for the propagation of trust and reputation, cooperation among selfish agents can be explained for a set of game theoretic simulations. For these simulations in particular, our proposal is shown to have provided more cooperative agent communities than existing schemes are able to.

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science

BRIEF CONTENT

CONTENT

ACKNOWLEDGEMENTS

First, I would like to acknowledge the Grace of the God, and for his unfailing faithfulness.

Then, I would like to thank my advisor Prof. Peter Szolovits for his kindness and patience with my years in his lab. Pete has guided my path in the past few years not only with his intellect and knowledge, but also with thoughtfulness about a young man's personal growth.

This work would not be possible without the guidance of my mentor and good friend Dr. Mojdeh Mohtashemi. This academic journey has been much more rewarding than otherwise because Mojdeh has taken upon herself to lead my path.

I also would like to acknowledge my other committee members: Prof. Isaac Kohane and Prof. Hal Abelson, for their help in various stages of my work. Zak in particular was the person who suggested that I should take up the study of reputation.

It is my great fortune to have met Dr. Jon Doyle in Pete's group. Jon is a true academic whose dedication to scholarship is rare even in MIT. I have greatly benefited from conversations with him about formalizing trust and reputation.

In the past few years, I have also recruited Ari Halberstadt to be a research partner in my journey through the simulation-land. His wisdom and experience in software design truly blows my mind. With his return to academics, I hope that we can continue our research partnership in the future.

I would not be a sane graduate student without a group of great friends. There are many whom I would like to thank: Jun Sun, Omar Bakr, Mike McGeachie, Dave Mitchell, Ying Zhang, Rosaria, Pamela Laree, and other members of MEDG. I would like to specially thank Fern DeOliveira for her help in getting my committee together and taking care of the administrative details for my defense. For my fun time with the Microsoft iCampus project, I am much indebted to the kindness of Jessica Strong, Becky Bisbee, and Dave Mitchell.

In my short time at Microsoft Research, I have greatly benefited from interactions and friendships with Dr. Marc Smith, Fernanda Viegas, Jeremy Goecks and other members of the Community Technologies Group.

I also would like to acknowledge the 3 master students whom I help guide their theses work the past few years: Cheewee Ang, Waikit Koh, and Francisco Tanudjaja. Their work has shortened the time that I remain a poor graduate student.

I am incredibly indebted to Marilyn Pierce, Peggy Carney, Merlene Ingraham and other staff in the EECS graduate office for their help in getting me over the hurdle of turning in a dissertation, and the financial support! I thank you all from the bottom of my heart.

For my short user study on citation analysis, I would like to thank people who have helped me with this study: Prof. Chris Haqq, Dr. Daniel Nigrin, Dr. Atul Butte and other endocrinologists at the Boston Children's Hospital.

Financially, I have been supported by various departments and agencies: Harvard/MIT Division of Health Science and Technology, MIT Department of Electrical Engineering and Computer Science, Whitaker Foundation, National Institute of Health/National Library of Medicine, Pfizer Corporation, Gillette Corporation, and Microsoft Corporation.

Most important of all, I would like to thank my family: mom, dad, and Stina, for your love, unwavering support and inspiration.

CHAPTER 1

Introduction

"Your corn is ripe today; mine will be so tomorrow. 'Tis profitable for us both that I shou'd labour with you today, and that you shou'd aid me tomorrow. I have no kindness for you, and know that you have as little for me. I will not, therefore, take any pains on your account; and should I labour with you on my account, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security."

-- David Hume, Treatise, III, II, section 5

"Covenants struck without the sword are but words."

-- Thomas Hobbes, Leviathan.

1.1 Trust and Reputation

In his *Treatise on Human Nature* (1737), David Hume provides a clear description on the problem involving trust. We rely on trust every day: we trust that our parents would support us, our friends would be kind to us, we trust that motorists on the road would follow traffic rules, we trust that the goods we buy have the quality commensurate with how much we pay for them, *etc.* Whether one believes societies to be results of divine ruling, or social contract, the very notion of society as an organizing body requires the notion of trust among its members.

Yet, underlying every social transaction is the temptation to defect against one's opponent so as to increase one's personal gain (or to lessen one's labor). For instance, after one party has received payment for the goods that he is selling, he is tempted to not deliver the goods or to provide goods that do not match the quality advertised to the purchasers. As illustrated by the opening quote of this chapter, parties engaged in many

social interactions are tempted to hold back cooperation even though doing so results in both being worse off. These interactions exhibit the structure of the well studied prisoner's dilemma (Axelrod, 1984).

Thomas Hobbes (1588-1679), an Enlightenment philosopher, argues that social dilemmas require formal mechanisms to guard individuals against defection. Such mechanisms must have some way to enforce any agreements or contracts so that sanctions are imposed on those who break them. Hobbes believes that the *State* should be the agency of enforcement, the "sword", for societies. Without the *State* and its laws, courts, and police, "the state of nature ... [would be in a] constant state of war, of one with all, ... the life of man, solitary, poor, nasty, brutish, and short." (Hobbes, *Leviathan*)

1.1.1 Formal Trust Producing Mechanisms

Formal mechanisms for tempering the temptation to defect refer to those that are institutionalized. Examples include legislations, written contracts, escrow services and the like. Disciplines such as law, finance and economics have been concerned with ensuring trust in societies through these formal mechanisms. In everyday activities such as credit card purchasing, many of us unknowingly have acquired trust for those whom we deal with through some formal trust producing systems. Institutions such as the credit report bureaus, the criminal justice system, loan collection agencies, *etc.*, have procedures and rules for guarding against potential defectors and to apprehend those who break the contracts or laws. Their existence guarantees to all parties involved in credit card transactions against defective activities.

1.1.2 Informal Trust Producing Mechanisms

Of course, not all social activities that we encounter every day are as well structured against defection as credit card purchasing. Further, the formal mechanisms for guaranteeing mutual cooperation are imperfect. Hence we must each take risks in trusting others to behave as we expect them (or as they promised) to do. We may use a variety of clues and past experiences to decide, in any particular transaction, whether to take such a risk. In these activities, we rely on indirect mechanism for guarding us against defection. These indirect mechanisms include very subjective factors such as our partners' firmness of handshakes, body language, accent, *etc.* We also rely on our social networks and the media to gather hearsay on others' experiences of dealing with these individuals. Given these information, we form opinions and infer the reputation about our potential partners. With reputation information, we then decide on how much we trust these individuals.

In addition to playing an information role, reputation also has a sanctioning role in social groups. If one violates others' trust, this person may be subject to loss of his or her good reputation. A tainted reputation can be very costly to one's future transactions.

Institutional guarantees such as legal contracts do not completely eliminate risks. Even when societies do provide adequate measures against these risks, they often involve lengthy steps that most people would prefer not to have to take. Nonetheless, in bilateral interactions involving risk, no cooperation would take place unless the party who moves first has a certain level of *trust* that the second party is willing to fulfill its obligations.

The study of trust outside formal mechanisms becomes more important in new communities where such mechanisms have yet to be firmly established. This is particularly the case for virtual (or electronic) communities today. These communities have created reputation or rating systems for the express purpose of encouraging trusting and trustworthy behaviors. The study of trust production in these communities is the subject of this dissertation.

1.2 Trust and Reputation in Virtual Communities

The past decade has seen the rise of virtual communities such as online chat rooms, electronic markets, and virtual multiplayer game worlds. These new spaces of human interaction have challenged our accumulated wisdom on how human interactions can occur. Many aspects of these virtual communities deserve to be and have been extensively researched (*e.g.*, Rheingold, 2000; Wellman, 2001; Smith and Kollock, 1999; Donath, 2002). In this dissertation we focus on one aspect that in many ways differs from the physical world equivalent: how trust and reputation are acquired and used.

An example virtual community is an electronic market such as eBay. Electronic markets hold great promises for bringing together large numbers of buyers and sellers across wide geographic regions on a scale that is unprecedented in the physical world. By so doing, electronic markets have the potential to dramatically improve the efficiency of trading through the reduction of search and transaction costs. For example, proponents of such markets claim that unlike the offline markets, buyers can choose the best possible deal for every transaction and work with different sellers every time. This claim hinges on an important assumption often unstated: that buyers and sellers can trust each other in ways that do not incur expensive transaction costs. For online trading communities, the production of trust is thus a crucial social factor that must be tackled.

Challenges facing trust in virtual communities arise ironically from what some proponents claim to be the advantages for such communities: being vast, nearly anonymous, simple to join and leave. By being vast,

- Members in virtual communities often span geopolitical boundaries where formal mechanisms ensuring trusts are difficult to establish.
- Virtual interactions lack direct physical cues such as tone of voice, body language, handshakes, store façade, *etc.*, which are often used as the first line for gauging trustworthiness in everyday interactions.
- Members are often anonymous and can enter and leave a community easily.
- Members often interact with strangers whom the members nor their friends have encountered before.[1]

The possibility for dealing with complete strangers without institutional guarantees significantly increases the risk for such interactions. Therefore, one expects trust to be hard to acquire. Hence, one would not expect to find many interactions requiring trust in such communities.

---

[1] In sociologists' terms, their "social networks" do not overlap (*c.f.*, Chapter 2 on sociology).

On the contrary, virtual communities have thrived in the case of eBay [2], internet newsgroups (Smith and Kollock, 1998), The WELL (Rheingold, 2000)[3], and ICQ [4], among others. In the case of eBay, trust between buyers and sellers is established with a set of simple rating schemes. What makes trust possible given the difficulties outlined above for virtual communities? The following tools contribute to trusting interactions online:

- **Escrow services**

  Some formal institutions already exist online for guaranteeing trust within individual countries. Services such as Paypal provide the equivalent of institutional guarantees for virtual communities. However, recent fraud cases involving Paypal point to some of the difficulties for such institutions that span geopolitical boundaries.[5]

- **History reporting for members**

  Virtual communities are capable of storing complete information on their members' interactions online. Companies such as eBay provide a breakdown of the number of positive, neutral, and negative ratings, and written feedbacks for a given seller over 1 week, 1 month and 6 months prior to any given transaction. Potential buyers can use this history reporting feature to evaluate their own risk profile for engaging in a transaction with the seller.

- **Reputation rating systems**

  A reputation or rating system attempts to provide succinct summaries of a user's history for a given virtual community. In the case of eBay (or Amazon), one's reputation is represented by a rating "score" which is calculated based on cumulative (or average ratings) by its members.

Similar to offline interactions, formal mechanisms offered by tools such as escrow services can only cover a small fraction of all interactions online. Except for financially related interactions, few institutional guarantees are available. For activities such as inferring experts, or finding compatible partners, or locating reliable opinions, more informal mechanisms are needed.

History and reputation reporting systems aim to provide the informal mechanisms for producing trust for online interactions. The majority of such systems today have been created by internet entrepreneurs and their properties have yet to be fully researched. Many are still primitive. The eBay reputation system is no more than a cumulative registry of user ratings and feedbacks on a given eBay member. Each feedback is accompanied by either a positive (+1), neutral (0), or negative (-1) rating. Clearly, human interactions are more finely grained and more sophisticated. In fact, one can easily think of schemes to take advantage of the eBay reputation system. Recent fraud

---

[2] http://www.ebay.com/
[3] http://www.well.com/
[4] http://www.icq.com/
[5] http://www.cnn.com/2002/TECH/industry/03/26/paypal.stranded.idg/index.html

cases on eBay remind us of the frailty of this simple design.[6] As virtual communities mature and grow to rely on these trust management systems, they deserve to be reexamined.

1.3 Contributions

This dissertation contributes to the study of trust and reputation by first providing a critical overview of the state of the art in this field. Many extant studies of trust and reputation have been made in the context of building reputation or rating systems for online communities. Most of these systems have been constructed without a formal rating model or much regard for our sociological understanding of these concepts.[7] Many such studies provide an intuitive approach to trust and reputation which appeal to common experiences without clarifying whether their usage of these concepts is similar to or different from those used by others.

We first describe a formal quantitative model for the rating process. Based on this model, we propose two personalized rating schemes and experimentally demonstrate their effectiveness for inferring trust using a simulated dataset and a real world movie-rating dataset. Our experiments show that the popular global rating scheme widely used in commercial electronic communities is inferior to our personalized rating schemes when sufficient ratings among members are available. The level of sufficiency is then discussed. In comparison with other models of reputation, we quantitatively show that our framework provides significantly better estimations of reputation. "Better" is discussed with respect to the rating process in Chapter 4 and then to two specific games to be discussed in Chapter 6 and Chapter 8.

One important contribution of this dissertation is the derivation of a mathematical framework for modeling trust and reputation that is rooted in findings from the social sciences. In particular, our framework makes explicit the importance of *social information* (*i.e.*, indirect channels of inference) in helping members of a social network choose whom they want to partner with or to avoid.

Finally, we extend our trust and reputation framework toward addressing a fundamental problem for social science and biology: evolution of cooperation. We show that by providing an indirect inference mechanism for the propagation of trust and reputation, cooperation among selfish agents can be explained. In a set of game theoretic simulations for evaluating the process for the evolution of cooperation, our proposal is shown to have provided more cooperative agent communities than many existing schemes are able to.

---

[6] *e.g.*, http://www.cnn.com/2000/TECH/computing/11/07/suing.ebay.idg/

[7] Examples of such systems include those in many commercial services such as Amazon or eBay, Zacharia and Maes (1999), Yu and Singh (2000), among others.

1.4 Relevance to Computer Science

Much of this dissertation is about sociological concepts such as society, trust and reputation. How do these concepts relate to contemporary computer science research?

As argued earlier in this chapter, the emergence of online communities opens new interaction spaces. In these virtual spaces, the advantages of face-to-face interactions, personal trust and reputation, and physical cues, among others, no longer apply. In this dissertation, we are specifically interested in the genesis and maintenance of reciprocity, trust and reputation in virtual communities. We believe that such social variables have significant roles for enhancing the user experiences online. As irrational as trusting and trustworthy behaviors might seem to selfishness-based modelers, such behaviors ultimately lead to advantages in other areas (such as survival of a social group). If these behaviors are so prevalent and important in our physical world, enabling such behaviors in the virtual worlds should pave the way to allow more people to be positively involved in these new interaction spaces than otherwise.

There has been extensive work for modeling cooperation, reciprocity, trust and reputation in diverse fields including computer science. Several chapters [8] in this work provide cross-disciplinary analyses of these concepts. This dissertation attempts to build on these diverse sources of scholarship to create new computational techniques for fostering cooperative behaviors in virtual communities.

A second reason why the discussion thus far should be relevant to computer science is that researchers in artificial intelligence have borrowed heavily from the machineries of economics in modeling rationality specifically and intelligence in general (Doyle, 1998; Wellman, 1993; Horvitz, 1986; Keeney and Raiffa, 1986; *etc.*). As economic modeling has so far neglected the modeling of norms, preferences, and other such social quantities (*c.f.*, Fehr, *et al.*, 2002), there exist many opportunities to explore and further artificial intelligence when one is equipped with deeper understandings of these concepts. Appendix C explores how this work is relevant to the modeling of (*ir*)rationality, cooperation and related social variables.

1.5 Roadmap

This chapter provides the motivating context for this work. Chapter 2 critically reviews the literature related to rating and reputation as a notion in particular. A typology is proposed to summarize this literature. Chapter 3 describes a formal rating framework for inferring and producing trust. We model the commercially popular global rating scheme within this rating framework. In this chapter, we also propose two personalized rating propagation schemes as alternatives to the global rating scheme. Chapter 4 presents simulations comparing our proposed rating systems with the global and a control schemes.

We examine the notion of trust and reputation more generally outside the rating problem in Chapter 5 and 6. Chapter 5 describes a sociologically justified, statistically sound computational model for trust and reputation. This model is based on the

[8] In particular, these chapters below provide these analyses: 2, 3, 5, 7 and 8.

personalized rating scheme proposed in Chapter 3 and experimentally tested in Chapter 4. With a set of evolutionary games known as iterated Prisoner's Dilemma, simulation results comparing our proposed computational model for trust and reputation with others are reported in Chapter 6.

Chapter 7 and 8 extend our trust and reputation framework and apply it to a fundamental problem in biology and social theories: how does cooperation evolve among self interested individuals? Several existing theories already attempt to answer this question; each with its merits and faults. We argue that cooperation can evolve among self-interested individuals if certain social structures are well-established. The social structures studied in this work are indirectly inferred trust and reputation. Chapter 7 is a review of the literature covering this field. Chapter 8 presents our computational framework for explaining the evolution of cooperation. The core of our proposal lies in modeling social interaction in the form of "social information" such as trust and reputation. We illustrate the robustness of our model through artificially simulated societies of agents. Finally, we compare our results to those obtained via other methodologies.

Chapter 9 briefly concludes this work and points to directions for future research opportunities.

CHAPTER 2

Notions of Reputation

Trust and reputation underlie almost every face-to-face trade. In an on-line setting, trading partners have limited information about each other's reliability or the product quality during the transaction. The analysis by Akerloff in 1970 on the Market for Lemons is also applicable to the electronic market. The main issue pointed out by Akerloff about such markets is the information asymmetry between the buyers and sellers. The buyers know about their own trading behavior and the quality of the products they are selling. On the other hand, the buyers can at best guess at what the sellers know from information gathered about them, such as their trustworthiness and reputation. Trading partners use each others' reputations to reduce this information asymmetry so as to facilitate trusting trading relationships.

Reputation or rating systems have become widespread for virtual communities. Such systems aim to enhance the level of trust among members, whether the goal is to increase number of auctions (*e.g.*, eBay), or to increase the sale of good products (*e.g.*, Amazon), or to engage in more social circles (*e.g.*, newsgroup). Before we investigate how such systems have been built and contribute to the production of trust for these communities, we review the research on reputation in this chapter, emphasizing the quantitative work that can be implemented in real world computer systems.

Reputation is not a single notion but one with multiple parts. Section 2.2 reviews the basic notions of reputation as used in several disciplines. Section 2.3 proposes a typology as a helpful framework to summarize existing notions of reputation. Chapter 6 describes a set of experiments and results aimed at understanding the relative strength of different notions of reputation as discussed in this chapter. A brief discussion of our typology concludes this paper.

2.1 Introduction

Reputation refers to a perception that an agent has of another's intentions and norms. Evolutionary biologists have used reputation to explain why selfish individuals cooperate (*e.g.*, Nowak and Sigmund, 1998). Economists have used reputation to explain "irrational" behavior of players in repeated economic games (*e.g.*, Kreps and Wilson,

1982). Computer scientists have used reputation to model the trustworthiness of individuals and firms in online marketplace (*e.g.*, Zacharia and Maes, 1999).

Reputation is often confused with concepts related to it, such as trust (e.g., Abdul-Rahman, *et al.*, 2000; Yu, *et al.*, 2001).[1] The trouble with a number of reputation studies lie in their lack of careful analysis based on existing social, biological, and computational literatures regarding reputation. We refer to Chapter 5 in this work, Ostrom (1998) or Mui, *et al.*, (2002) for a clarification of reputation, trust, and related concepts.

2.2 Background

Trust and reputation are concepts studied by researchers and thinkers in different fields. The sections below by no mean attempt to divide these works into distinct buckets. Rather, they group together similar works that have often been of interest to audiences of the various disciplines.

2.2.1 Reputation Reporting System

Reputation reporting systems have been implemented in e-commerce systems and have been credited with these systems' successes (Resnick, *et al.*, 2000a). Several research reports have found that seller reputation has significant influences on on-line auction prices, especially for high-valued items (Houser and Wooders, 2000; Dewan and Hsu, 2001).

The reputation system in eBay is well studied. *Reputation* in eBay is a function of the cumulative positive and non-positive ratings for a seller or buyer over several recent periods (week, month, 6-months). Resnick and Zeckhauser (2000b) have empirically analyzed this reputation system and conclude that the system does seem to encourage transactions. Houser and Wooders (2000) have used games to study auctions in eBay and describe reputations as the *propensities to default* – for a buyer, it is the probability that if the buyer wins, he will deliver the payment as promised before the close of the auction; for a seller, it is the probability that once payment is received, he will deliver the item auctioned. Their economic analysis shows that reputation has a statistically significant effect on price. Both Lucking-Reily, *et al.* (1999) and Bajari and Hortacsu (2000) have empirically examined coin auctions in eBay. These economic studies have provided empirical confirmation of reputation effects in internet auctions.

Despite the obvious usefulness of reputation and related concepts for online trading, conceptual gaps exist in current models about them. Resnick and Zeckhauser (2000b) have pointed out the so called *Pollyanna* effect in their study of the eBay reputation reporting system. This effect refers to the disproportionately positive feedbacks from users and rare negative feedbacks. They have also pointed out that despite the incentives to free ride (for not providing feedbacks), feedbacks by agents are provided in more than half of the transactions. This violates the rational alternative of

---

[1] The relationship between trust and reputation (and other related social quantities) is described in the next chapter.

taking advantage of the system without spending the effort to provide feedback. Moreover, these studies do not model deception and distrust. As shown by Dellarocas (2000), several easy attacks on reputation systems can be staged. These studies also do not examine issues related to the ease of changing one's pseudonym online. As Friedman and Resnick (1998) have pointed out, an easily modified pseudonym system creates the incentive to misbehave without paying reputational consequences.

2.2.2 Economics

Economists have extensively studied reputation in game theoretic settings. Many of the economic studies on reputation relate to repeated games. In particular, the Prisoner's Dilemma or the Chain Store stage game is often used in these studies (*e.g.*, Andreoni and Miller, 1993; Selten, 1978). In such repeated games, reputation of players is linked to the existence of cooperative equilibria. Game theorists have postulated the existence of such an equilibrium since the 1950's in the so called *Folk Theorem* (Fudenberg and Maskin, 1986). However, the first proof did not come until 1971 in the form of discounted *publicly observable* repeated games between two players (Friedman, 1971). Recent development in game theory have extended this existence result to *imperfect publicly monitored* games and to some extent *privately monitored* games (Kandori, 2002), and to games involving changing partners (Okuno-Fujiwara and Postelwaite, 1995; Kandori, 1992). Economists often interpret the sustenance of cooperation between two players as evidence of "reputation effects" (Fudenberg and Tirole, 1991).

Entry deterrence is often studied by game theorists by using notions of reputation. Kreps and Wilson (1982) borrows Harsanyi (1967)'s theory of imperfect information about players' payoffs to explain "reputation effects" for multi-stage games involving an incumbent firm versus multiple new entrants. They show that equilibria for the repeated game exist (with sufficient discounting) so that an incumbent firm has the incentive to acquire an early reputation for being "tough" in order to decrease the probability for future entries into the industry. Milgrom and Roberts (1982) report similar findings by using asymmetric information to explain the reputation phenomenon. For an incumbent firm, it is rational to seek a "predation" strategy for early entrants even if "it is costly when viewed in isolation, because it yields a reputation which deters other entrants." (*ibid.*) More recently, Tirole (1998) and Tadelis (2000a) have studied reputation at the firm level — firm reputation being a function of the reputation of the individual employees. Tadelis (2000b) has further studied reputation as a tradeable asset, such as the tradename of a firm.

2.2.3 Scientometrics

Scientometrics (or bibliometrics) is the study of measuring research outputs such as journal impact factors. Reputation as used by this community usually refers to number of cross citations that a given author or journal has accumulated over a period of time (Garfield, 1955; Baumgartner, *et al.*, 2000). As pointed out by Makino, *et al.*, 1998 and

others, cross citation is a reasonable but sometimes confounded measure of one's reputation.

2.2.4 Computer Science

Trust between buyers and sellers can be inferred from the reputation that agents have in the system. How this inference is performed is often hand-waved by those designing and analyzing such systems as Zacharia and Maes (1999), Yu and Singh (2000), Houser and Wooders (2001). As briefly discussed earlier, several easy attacks on reputation systems can be staged (Dellarocas, 2000).

Besides electronic markets, trust and reputation play important roles in distributed systems in general. For example, a trust model features prominently in Zimmermann's Pretty Good Privacy system (Zimmermann, 1995; Khare and Rifkin, 1997). The reputation system in the anonymous storage system Free Haven is used to create an accountability system for users (Dingledine, *et al,* 2001). Trust management in the system Publius allows users to publish materials anonymously such that censorship of and tampering with any publication in the system is rendered very difficult (Waldman, *et al.*, 2000).

In the computer science literature, Marsh (1994) is among the first to introduce a computational model for trust in the distributed artificial intelligence (DAI) community. He did not model reputation in his work. As he has pointed out, several limitations exist for his simple trust model. Firstly, trust is represented in his model as a subjective real number between the arbitrary range –1 and +1. The model exhibits problems at the extreme values and at 0. Secondly, the operators and algebra for manipulating trust values are limited and have trouble dealing with negative trust values. Marsh also pointed to difficulties with the concept of "negative" trust and its propagation.

Abdul-Rahman, *et al,* (2000) have studied reputation as a form of social control in the context of trust propagation — reputation is used to influence agents to cooperate for fear of gaining a bad reputation. Although not explicitly described, they have considered reputation as a propagated notion which is passed to other agents "by means of word-of-mouth".

Sabater, *et al.* (2001) have defined reputation as the "opinion or view of one about something" and have modeled 3 notions of reputation: individual, social, and ontological. Individual reputation refers to how a single individual's impressions are judged by others. Social reputation refers to impression about individuals based on the reputation of the social group they belong to. Ontological refers to the multifaceted nature of reputation – depending on the specific context.

Mui, *et al.,* (2001) and Yu, *et al.*, (2001) have proposed probabilistic models for reputation. The former uses Bayesian statistics while the latter uses Dempster Shafer evidence theory. Reputation for an agent is inferred in both cases based on propagated ratings from an evaluating agent's neighbors. These propagated ratings are in turn weighted by the reputation of the neighbors themselves.

2.2.5 Evolutionary Biology

A detailed review of how trust and reputation are relevant for evolutionary biologists is given in Chapter 7. Here we briefly highlight the more recent research of relevance to the next few chapters.

Among evolutionary biologists, Pollock and Dugatkin (1992) have studied reputation in the context of iterated prisoners' dilemma games (Axelrod, 1982). They have introduced a new interaction strategy (named *Observer Tit For Tat*) which determines whether to cooperate or defect based on the opponent's reputation. Reputation here is inferred from the ratio of cooperation over defection. Nowak and Sigmund (1998, 2000) use the term *image* to denote the total points gained by a player by reciprocation. The implication is that image is equal to reputation. Image score is accumulated (or decremented) in every direct interaction among agents. Following the studies by Pollock and Dugatkin (1992), Nowak and Sigmund (1998) have also studied the effects of third party observers of interactions on image scores. Observers have a positive effect on the development of cooperation by facilitating the propagation of observed behavior (image) across a population. Castelfranchi, *et al.* (1998) explicitly have reported that communication about "Cheaters"' bad reputations in a simulated society is vital to the fitness of agents who prefer to cooperate with others.

2.2.6 Anthropology

Anthropologists describe the observed human cooperation as "altruistic" since selfishness-based arguments cannot explain such behaviors (Ensminger, 2002; Henrich, *et al.*, 2002).

Socio-biological theories predict that cooperation and altruism should be limited to kin and reciprocating partners (Hamilton, 1963; Trivers, 1971; Axelrod, 1984; Boyd and Richerson, 1989). However, humans cooperate with large groups of unrelated individuals who do not promise reciprocation. Their cooperation is not just co-incidental to their selfish pursuit; anthropological experiments with western subjects have shown that these individuals actually have *social preferences* that support large scale cooperation (Fehr, *et al.*, 2001). Such preferences include: inequality aversion, strong reciprocity, and concerns for fairness.

One of the main goals of the recently completed MacArthur Cross-Cultural Project is to answer whether the canonical selfishness-based models of human decision making holds true across 18 distinct social-economic groups in 4 continents with over 1030 subjects (Henrich, *et al.*, 2002). The results by 12 researchers in economics and anthropology emphatically show that the canonical selfishness-based assumption about human do not explain any of the social groups studied. At the same time, behavioral variability in this well-designed and controlled set of experiments point to a lack of universal pan-human explanation for issues about cooperation and related variables such as reciprocity and trust.

These recent anthropological results have greatly challenged our understanding on concepts such as cooperation, rationality, reputation, *etc.* As pointed out by Chapter 9 of

this work, incorporating these new findings is a major area of future extension in our work.

2.2.7 Sociology

Reputation is of interested to sociologists for the obvious reason that reputation is a social phenomenon. The rise of large scale virtual communities as new spaces of human interaction has challenged sociologists' accumulated wisdom on how human interactions occur (*e.g.*, Smith and Kollock, 1999; Rheingold, 2000; Wellman, 2001; Donath, 2002). Research of reputation in newsgroups, rating systems, and mediated spaces is given new prominence (*e.g.*, Kollock and Smith, 1996; Resnick and Zeckhauser, 2000b; Rheingold, 2000).

Among the *structural* sociologists studying social networks, reputation is often studied as a network parameter associated with a society of agents (Freeman, 1979; Krackhardt, *et al.*, 1993; Wasserman and Faust, 1994). Reputation or prestige is often measured by various centrality measures. An example is a measure proposed by Katz (1953) based on a stochastic coincidence matrix where entries record social linkages among agents. Because the matrix is stochastic, the right eigenvector associated with the eigenvalue of 1 is the stationary distribution associated with the stochastic matrix (Strang, 1988). The values in the eigenvector represent the reputation (or *prestige*) of the individuals in the society. Unfortunately, each individual is often modeled with only one score, lacking context dependence.

Other sociologists and social scientists who study reputation from a qualitative perspective are interested in the social context surrounding reputation (or "status", "prestige", "power", "dominance", *etc.*). Types of reputation can be divided into "ascribed" (*e.g.*, chiefdoms and states), "achieved" (Renfrew & Bahn, 1996), or "earned" (*e.g.*, by excelling at specific activities), "forced" (through threat, fear, persuasion or compulsion, Krackle, 1978).

In her Presidential Speech to the American Political Science Society, Ostrom (1998) has argued for a holistic approach to study reputation based on how reputation, trust, and reciprocity interrelate. Based on her qualitative model, a computational model for these related concepts has been proposed by Mui, *et al.* (2002).

2.3 Reputation Typology

2.3.1 Contextualization

Reputation is clearly a context-dependent quantity. For example, one's reputation as a computer scientist should have no influence on his or her reputation as cook. Formal models for context-dependent reputation have been proposed by Mui, *et al.*, (2001) and Sabater, *et al.*, (2001), among others. Existing commercial reputation systems in eBay or Amazon provide only one reputation rating per trader or per book reviewer. Context-dependent reputation systems (*e.g.*, based on value of items) might help mitigate

cybercrimes involving self-rating on small value items among a cartel of users for gaining reputation points (*c.f.*, US Dept of Justice, 2001).

2.3.2 Personalization

Reputation can be viewed as a *global* or *personalized* quantity. For social network researchers (Katz, 1953; Freeman, 1979; Marsden, *et al.*, 1982; Krackhardt, *et al.*, 1993), prestige or reputation is a quantity derived from the underlying social network. An agent's reputation is globally visible to all agents in a social network. In the same way, scientometricians who use citation analysis to measure journal or author impact factors (*i.e.*, reputation) also rely on the underlying network formed by the cross citations among the articles studied (Garfield, 1955; Baumgartner, *et al.*, 2000). Many reputation systems rely on global reputation. In the case of Amazon or eBay, reputation is a function of the cumulative ratings on users by others. Global reputation is often assumed in research systems such as those in Zacharia and Maes (1999)'s Sporas, Nowak and Sigmund (1998)'s image score without observers, Rouchier, *et al.* (2001)'s gift exchange system, among others.

Personalized reputation has been studied by Zacharia and Maes (1999), Sabater, *et al.*, (2001), Yu, *et al.* (2001), among others. As argued by Mui, *et al.* (2002), an agent is likely to have different reputations in the eyes of others, relative to the *embedded social network*. The argument is based on sociological studies of human behavior (*c.f.*, Granovetter, 1985; Raub and Weesie, 1990; C. Castelfranchi, *et al.*, 1998). Depending on factors such as environmental uncertainties, an agent's reputation in the same embedded social network often varies (Kollock, 1994).

How many notions of reputation have been studied? Based on the reviewed literature, an intuitive typology of reputation is proposed as shown in Figure 2.1. This typology tree is to be discussed one level at a time in the rest of this section. Each sub-section reviews reputation literatures that are relevant to that part of the tree.

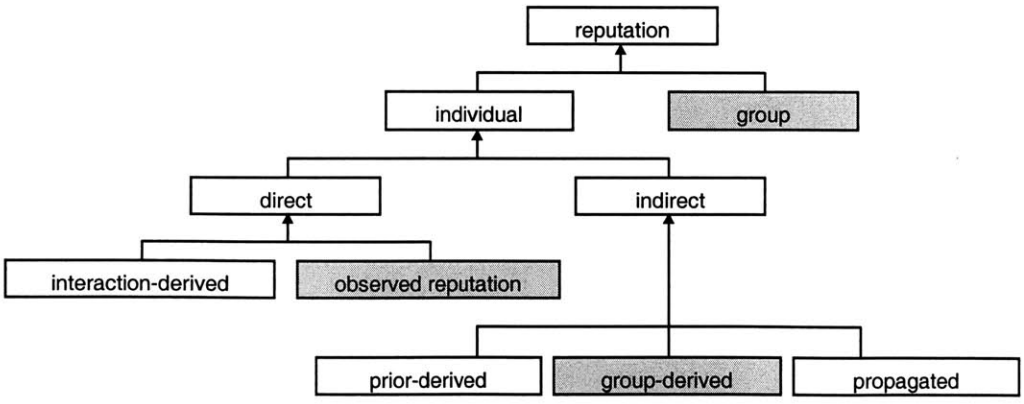
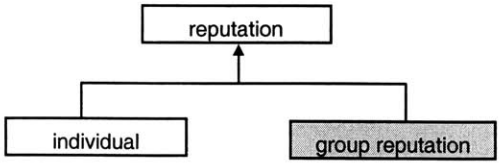


Figure 2.1 *Reputation typology. It is assumed that reputation is context dependent. Shaded boxes indicate notions that are likely to be modeled as social (or "global") reputation as opposed to being personalized* to the inquiring agent (see text).

2.3.3 Individual and Group Reputation

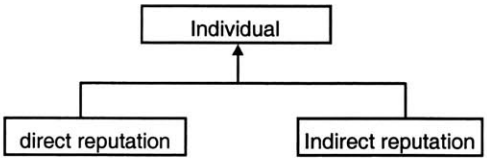


At the topmost level, reputation can be used to describe an individual or a group of individuals. Existing reputation systems such as those in eBay, Amazon, Free Haven, or Slashdot (*c.f.*, Resnick, *et al.* 2000b; Houser and Wooders, 2001; Dingledine, *et al.*, 2001) concentrate on reputation of the individuals.

Economists have studied group reputation from the perspective of the firm (Kreps and Wilson, 1982; Tirole, 1996; Tadelis, 2000). A firm's (group) reputation can be modeled as the average of all its members' individual reputation. Among computer scientists, Sabater and Sierra (2001) have studied the *social* dimension of reputation, which is inferred from a group reputation in their model. Halberstadt and Mui (2001) have proposed a hierarchical group model and have studied group reputation based on simulations using the hierarchical model. Their group model allows agents to belong to multiple overlapping groups and permits reputation inferences across group memberships.

Commercial groups such as Reputation.com and OpenRatings [2] are applying their proprietary algorithms to manage buyer-supplier company relationships based on individual transactions. Inherent in these models is the distinction between individual and group reputation.

2.3.4 Direct and Indirect (individual) Reputation

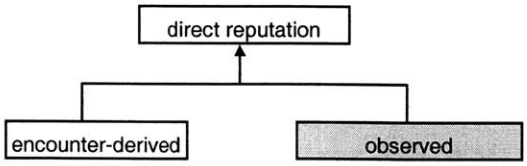


One can consider individual reputation as derived either (1) from direct encounters or observations or (2) from inferences based on information gathered indirectly. Direct reputation refers to reputation estimates by an evaluator based on direct experiences (seen or experienced by the evaluating agent first hand). Indirect reputation refers to reputation estimates that are based on second-hand evidence (such as by word-of-mouth).

---

[2] *c.f.*, http://www.reputation.com and http://www.openratings.com

2.3.5 Direct Reputation



Direct experience with another agent can be further divided into (1) observations made about another agent's encounters with others, and (2) direct experience interacting with that other agent.

2.3.5.1 Observed Reputation

Reputation rating in systems such as eBay provides an example for both observed and encounter-derived reputation. These ratings are direct feedbacks from users about others with whom they have interacted directly. After an encounter with a seller, a buyer can provide a rating feedback which can directly affect a seller's reputation in the system — this is *encounter-derived* reputation (Dewan and Hsu, 2001; Resnick and Zeckhauser, 2000b). Buyers who have not interacted with a seller need to rely on others' ratings as observations about a seller — thereby deriving *observed reputation* about the seller.

Observer based reputation plays an important role in reputation studies by evolutionary game theorists and biologists. Pollock and Dugukin (1992) have introduced "observed tit-for-tat" (OTFT) as an evolutionarily superior strategy compared to the classic tit-for-tat strategy for the iterated Prisoner's Dilemma game. OTFT agents observe the proportion of cooperation of other agents. Based on whether a cooperation threshold is reached, an OTFT agent determines whether to cooperate or defect on an encounter with another agent. Similarly, Nowak and Sigmund (1998) use observer agents to determine agent actions in their image-score based game.

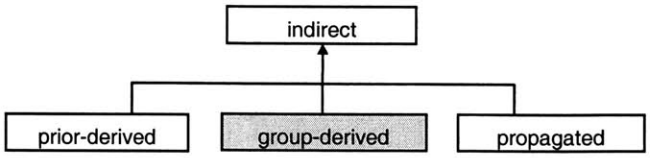
2.3.5.2 Encounter-derived Reputation

In our terminology, "observed" reputation differs from "encounter-derived" reputation in that the latter is based on actual encounters between a reputed agent and his or her evaluating agent. For example, journal impact factor as determined by citation analysis (Garfield, 1955) is an "observed" reputation based on the observed cross-citation patterns.[3] However, individual researchers might not agree with the impact factor based

[3] Anthropomorphically, each journal article's citation is a "rating feedback" to the cross-citation analysis observer.

on their own readings of individual journals.[4] Each researcher revises the observed reputation based on their direct experience with each journal. Field studies by Kollock (1994) have shown that personal interactions play a more important role than indirect observations in determining whether users choose to interact with one another socially.[5]

2.3.6 Indirect Reputations



Without direct evidence, individual reputation can be inferred based on information gathered indirectly.

2.3.6.1 Prior-derived reputation

In the simplest inference, agents bring with them prior beliefs about strangers. In human societies, each of us probably has different prior beliefs about the trustworthiness of strangers we meet. Sexual or racial discrimination might be a consequence of such prior beliefs.

For agent systems, such discriminatory priors have not yet been modeled. Mui, *et al.*, (2001)'s probabilistic model uses a uniform distribution for reputation priors. This is equivalent to an ignorance assumption about all unknown agents. Zacharia and Maes (1999)'s system give new agents the lowest possible reputation value so that there is no incentive to throw away a cyber identity when an agent's reputation falls below a starting point. Nowak and Sigmund (1998)'s agents assume neither good nor bad reputation for unknown agents.

2.3.6.2 Group-derived Reputation

Models for groups can been extended to provide prior reputation estimates for agents in social groups. Tadelis (2001)'s study of the relation between firm reputation and employee reputation naturally suggests a prior estimate based on the firm that an economic agent belongs to. If the firm has a good reputation, the employee can benefit from being treated as if he or she had a good reputation, and vice versa. In the computer science field, both Sabater and Sierra (2001), and Halberstadt and Mui (2001) have postulated different mapping between the initial individual reputation of a stranger and the group from which he or she comes from. Since the reputation of a group can be

---

[4] Citation analysis based impact factor has been questioned on scientific ground (Makino, *et al.*, 1998).

[5] Our term "Encounter-derived" reputation is usually called "personalized" (Zacharia and Maes, 1999; Sabater and Sierra, 2001; Yu and Singh, 20001; Mui, *et al.*, 2001). We avoid the word "personalized" here since other notions of reputation in Figure 2.1 can also be described as such.

different to different agents, individual reputation derived from group reputation is necessarily personalized to the evaluating agent's perception of the group.

2.3.6.3 Propagated Reputation

Finally, although an agent might be a stranger to the evaluating agent, the evaluating agent can attempt to estimate the stranger's reputation based on information garnered from others in the environment. As Abdul-Rahman and Hailes (2000) have suggested, this mechanism is similar to the "word-of-mouth" propagation of information for humans. Reputation information can be passed from agent to agent. Schillo, *et al.*, (2000), Mui, *et al.*, (2001) Sabater and Cierra, (2001), and Yu and Singh (2001) have all used this notion, that reputation values can be transmitted from one agent to another. What differentiates these approaches is the care taken in combining the information gathered from these chains. Yu and Singh (2001) have tried to use Dempster-Shafer theory for this combination. Mui, *et al.*, (2001) have used Bayesian probability theory. The latter has also used the Chernoff Bound to propose a reliability measure for information gathered along each chain.

2.4 Discussions

This chapter has proposed a typology for different notions of reputation that have been studied by various researchers and implemented in real world systems. The typology serves a useful function in unifying the diverse literature on reputation. Based on this typology, Chapter 6 will experimentally study the relative strengths of different notions of reputation in a set of evolutionary games. Whereas these notions of reputation could only be compared qualitatively before, our experimental framework will enable us to compare them quantitatively.

Reputation has become a popular topic for building online rating systems (*e.g.*, Sycara, *et al.*, 1999; Zacharia and Maes, 1999; Yu and Singh, 2000; Dingledine, *et al.*, 2001). As mentioned earlier in Chapter 1, many of these models have been constructed without a formal rating model or much regard to our sociological understandings of these concepts.

The next chapter formalizes a reputation or rating model for inferring trust. It also proposes two personalized rating systems using this rating model. Chapter 4 reports experiments on the proposed rating systems. Chapter 5 and 6 discuss trust and reputation with a sociological perspective.

CHAPTER 3

Online Rating Systems

In our everyday lives, we have various opinions about whether we approve of one another in different situations. For example, we might approve or disapprove of George W. Bush in his foreign policy toward Iraq; we might approve or disapprove our partners' outfit for a concert. The level of approval varies. We could be strongly against, for, or ambivalent about one another for a given situation.

Based on our conceived level of approval for one another, we form trust relationships over time. The sociological aggregate of individuals' opinions of one another can be interpreted as the basis of individuals' "reputation" in a society. This chapter examines schemes for how approval of one toward another can be propagated via ratings about other members of a community.

For distributed systems at large and e-commerce systems in particular, ratings play an increasingly important role. Ratings confer reliability or reputation measures about sources. This chapter reports our formalization of the rating process. We argue that ratings should be context- and individual-dependent quantities. In contrast to existing rating systems in many e-commerce or developer sites, our approach makes use of personalized and contextualized ratings for assessing source reliability and reputation. We present two new approaches to estimating reputation from ratings based on indirect inference. In the next chapter, we will report experimental results about our proposed system. Through the formalism and the experiments, we aim to show that indirect inference using personalized and contextualized ratings can enhance the production of trust, much more so than many existing methods based on non-personalized ratings.

3.1 Rating Systems: Trust and Reputation Inference

As alluded to in Chapter 1, philosophers have argued that societies depend on trust being present (Baier, 1986). Under this premise, virtual communities have much going against them. These communities cover vast geographic distances; they are easy to join and leave; their members are semi-anonymous.

By being vast, virtual communities span geopolitical boundaries where formal mechanisms (*e.g.*, law) ensuring trust are difficult to implement. Because it is easy for members to join and leave, persistent identities, which are important for trust, become difficult to establish. Even the identities created online are *semi-anonymous* in the sense

that these identities correspond mainly to email addresses, avatars, or credit card numbers; each real world individual can easily create multiple identities. Anonymity is good for protecting individual privacy but is a hindrance to the production of trust by others.

The fact that many virtual communities have been highly successful is due to their rating systems. There are many types of rating systems, with ratings of people, content, objects, *etc.* For the convenience of discussion below, we consider a generic rating system for an auction web site, such as eBay. The ratings concern individual buyers and sellers.

Under such a generic rating system, the collection of ratings of a given user represents the reputation of that member, as rated by other members in the community. Using this reputation measure, for example, a buyer can determine whether the seller is a party that could be trusted to carry out a transaction successfully. If that were not the case, the buyer could use the system to look for a seller that is more trustworthy.

One useful view of a rating system is through the lens of a social network (Wasserman and Faust, 1994). A social network is a directed graph often used by sociologists to represent a community. In a social network, members of a community can be represented by nodes, and the edges between the nodes represent the existence of ratings between users. The weight associated with each edge is the current rating that is given by one user to another based on their past and current interactions. The rating system updates these weights as the ratings change over time. The reputation of a member is then calculated depending on the structure of the network as well as the ratings between the members. For an example of a social network, see Figure 3.3.

3.2 Rating Systems: Background

In the abstract, a rating system is the intermediary that allows ratings of objects to propagate from one user to another. In the commercial world, a number of rating systems have been built; but most of these are built in an *ad-hoc* fashion. These systems usually assign a single reputation to each user and therefore ignore the personalized nature of reputation. An individual's reputation, as opposed to being a fixed attribute, actually varies depending on the tastes, preferences, opinions, and biases of other people interested in knowing his reputation. We term the assignment of reputation that is not personalized as "global" reputation. Dellarocas (2000) has provided warnings against possible attack methodologies that can be used against ratings management systems that employ a "global" notion of reputation.

As an example of such commercial systems is the eBay rating system is a cumulative registry of user feedbacks on a given eBay member. Each feedback is accompanied by either a positive (+1), neutral (0), or negative (-1) rating. Clearly, human interactions are more finely grained and more sophisticated. In fact, one can easily think of schemes to take advantage of the eBay reputation system, as discussed in the last chapter.[1]

---

[1] Also see news article such as http://www.cnn.com/2000/TECH/computing/11/07/suing.ebay.idg/

Sporas and Histos (Zacharia and Maes, 1999) allow each user in a community to rate one another after each transaction and modify their reputations based on these ratings. Reputation of a user in Sporas is determined *globally* in a similar way as that in eBay, based on the average of all ratings given to an agent. Histos is a *personalized* rating system with each inquirer receiving ratings about others based on who makes a query and the local environment surrounding the inquirer. Unfortunately, the formulation of Histos is based on informal arguments about what one expects of a *reasonable* rating system.

In recent literature, there have been many attempts to formulate a coherent set of rules for designing systems that manage reputation ratings. Among these, Glass and Grosz (2000) have proposed a "brownie-points" system to represent how conscientious an agent is in a community. Yu and Singh (2000), and Rouchier *et al.* (2001) have each suggested different formulae for calculating reputation ratings among agents. Many of these existing schemes are based on the "global" notion of reputation.

Noted exceptions are attempts that use collaborative filtering techniques. Collaborative filtering allows similarity between users to be estimated (Resnick, *et al.*, 1994; Goldberg, *et al.*, 1992). Weighting ratings or reputation measures with the similarity score calculated from collaborative filtering offers one technique for personalizing ratings (Lashkari, *et al.* 1994). This technique stems from mathematical considerations of metric space and no one has yet established its equivalent in real social settings. In addition, collaborative filtering techniques suffer from two related problems:

- Slow initialization: after a collaborative filtering system is set up, similarity among users cannot be estimated reliably due to lack of user data.
- Large user base requirement: to provide reasonable similarity measures, the number of sample points (user preferences) must be sufficiently large.

Other systems such as Platform for Internet Content Selection (PICS) [2], Better Business Bureau, Weaving a Web of Trust (Khare and Rifkin, 1997), and Yenta (Foner, 1997), users are required to assert their own reputation rating of themselves, and either have authoritative agencies or other users verify their assertions.

[2] http://www.w3.org/PICS/

3.3 Formalizing the Rating Process

This section first presents a formal model for the rating process. How an agent update his or her belief about the trustworthiness of others can be considered a learning process. An abstract learning model is proposed. A discussion then follows on extending the rating model to indirect inferences based on ratings of members in a social network. The notion of a rating propagation function is introduced that forms the basis for our proposals for estimating trust among members in a community.

3.3.1 The Rating Model

To formalize the rating process, the following model is used:

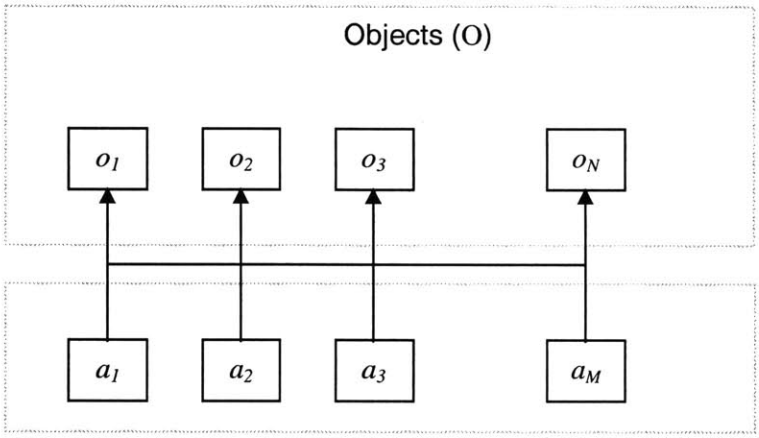


Figure 3.1 Model for the abstract rating process, where arrows indicate "ratings" by agents in A for objects in O.

Agents exist in an environment of objects (O) and other agents:

Set of agents: $A = \{ a_1, a_2, \ldots a_M \}$ (3.1)

Set of objects: $O = \{ o_1, o_2, \ldots o_N \}$ (3.2)

where objects can represent agents. When an agent would like to rate other agents, the other agents belong to the set O.

In this model, only 2 ratings by an agent are considered: an "approve" (represented by '1') or "disapprove" (represented by '0') for an object o_k in the environment. Let this rating process be represented by:

Rating: $\rho : A \times O \rightarrow \{ 1, 0 \}$ (3.3)

where ρ_{ik} represents the rating by agent a_i on object o_k.

To model the process of opinion sharing between agents, the concept of an encounter is required. An encounter is an event between 2 different agents (a_i, a_j) such that the query agent a_i asks the response agent a_j for a_j's rating of an object:

Encounter: $e \in E = A^2 \times O \times \{ 0,1 \} \cup \{ \perp \}$ (3.4)

Two cases are next considered for sharing opinions among agents in A. The first case below takes all objects in O to be in the same "context" (*i.e.*, all objects in O are related to the same subject matter). The second case considers multiple contexts. Clearly, sharing approval ratings in multiple contexts need to take into account the similarities and differences among the various contexts.

3.3.1.1 Uniform Context Environment

Consider an agent a_i in the model shown in Figure 3.1 who has never interacted with object o_k in the past. Before taking the trouble of interacting with o_k, a_i asks other agents (A_{-i}) in the environment what their ratings for o_k are. a_i will decide to interact with object o_k if the weighted sum of the ratings from agents in A_{-i} is greater than a certain threshold $thres_i$. Determination of $thres_i$ is a system policy issue not discussed here.

The weights on the ratings from other agents are determined by a_i's level of approval of other agents about the objects in the (uniform context) environment. The higher the approval a_j has in a_i's mind, the higher the weight a_i gives to a_j's rating of object o_k.

Reputation of a_j in a_i's mind can be considered as the probability that in the next encounter, a_j's rating about a new object in a given context will be approved by a_i, each given independently. The reputation probability can be represented by the mapping:

$$\text{Reputation:} \qquad R : \mathrm{A} \times \mathrm{A} \rightarrow [\,0,1\,] \tag{3.5}$$

where a_i's approval of itself can be defined as 1. No object is mapped from the domain side of this mapping since the mapping is about *any generic* object in the context of interest not yet encountered. This mapping is the subject of the next section in this paper.

The state of the system is the set of reputations:

$$\text{State:} \qquad \mathrm{S} = \mathrm{R} \tag{3.6}$$

The history of the system can be represented by:

$$\text{Set of Times:} \qquad \mathrm{T} = \{\, 0, 1, \ldots, t \,\} \tag{3.7}$$

$$\text{History:} \qquad h \in \mathrm{H} \quad \textit{and} \quad h : \mathrm{T} \rightarrow \mathrm{S} \times \mathrm{E} \tag{3.8}$$

3.3.1.2 Multiple Contexts Environment

Ratings can be shared across encounters easily if the environment has a single context. How should ratings be evaluated when there are multiple contexts?

First, a context has to be defined. A context is a set of attributes and their instantiated values about an environment. Let an attribute be defined as the presense ('1') or absense ('0') of a trait. The set of all attributes is possibly countably infinite, and is defined as follows:

$$\text{Attribute:} \qquad b \in \mathrm{B} \quad \textit{and} \quad b : \mathrm{O} \rightarrow \{\, 0, 1 \,\} \tag{3.9}$$

$$\text{Set of attributes:} \qquad \mathrm{B} = \{\, b_1, b_2, \ldots \,\}$$

A context is then an ordered list of instantiated attributes:

Context: $c = < b_i', b_j', \ldots b_k' > \quad where\ c \in C$ (3.10)

where each element of c is an instantiated value for the corresponding attribute. We assume that there are no duplicate attributes in the list.

In a multiple-contexts environment, any agent's reputation is clearly context-dependent. The reputation mapping can now be represented by:

Reputation: $R : A \times A \times C \rightarrow [0, 1]$ (3.11)

Therefore, for an encounter i, the binary random variable $x_{ab}(i)$ represents a's approval of b after the i^{th} encounter between them.

3.3.2 Multi-context Reputation

Inference in a multi-context environment involves ontological techniques that can form a separate science in itself. To avoid diverting attention from the rating process here, we refer interested readers to the paper by Koh and Mui (2001).

Transference of one's reputation from one context to the next is commonly applied in our everyday activity. For example, most people can infer that an agent i's reputation as a good politician is likely to mean that i is probably a good public speaker also. However i's reputation as a cook cannot be inferred from i's reputation as a politician. The context of being a politician is likely to share more instantiated attributes for being a public speaker than for being a cook. Some of these attributes include: *being confident, being eloquent, knowing the key issues of the day, etc.* To properly measure the level of transference of one's reputation from one context to the next is outside the scope of this work. For simplification, we consider the most stringent case of no reputation transference from one context to the next in this work. In other words, one's reputation in one context has no effect on his or her reputation in another context.

3.3.3 Reputation Learning

Reputation for an agent by others is learned over encounters among them. Assume that reputation only changes after an actual encounter (either directly with the individual involved, or through reputation sharing among other agents). After each encounter, how is the update of agent a_i's reputation for a_j?

The update rule for the state (S) of the system can be represented as:

Update rule: $newstate : S \times E \rightarrow S$ (3.12)

3.3.4 Indirect Inference by Rating Propagation

We represent an online community using a social network where nodes represent members and directed edges represent direct pair-wise ratings. Each social network is formed with respect to a specific context. By indirect inference, we refer to the ability to estimate the rating that a subject would have given to another member in the community as if that subject has directly interacted with that member.

In any sizeable social networks, members are unlikely to have interacted with (rated) every other members. Humans rely on gossip or word-of-mouth to propagate

their opinions about each other. In evaluating a stranger's trustworthiness, we weight those of our friends' opinions about this stranger by how much we trust our friends and come to a conclusion on whether we are going to trust this stranger. Hence, propagation of opinions (of which ratings is one) in human society is a very natural phenomenon.

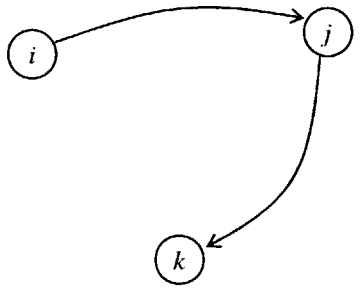


Figure 3.2. Illustration for indirect inference of i's rating for k based on i's rating on j and j's rating on k.

Let $\rho_{ij}(c)$ be the rating that member i gives to member j with respect to context c. Assume that $\rho_{ij}(c)$ represents all the information that i has about j. Given $\rho_{ij}(c)$ and $\rho_{jk}(c)$ where $i \neq j \neq k$, how i should evaluate k can be expressed as:

$$\rho_{ik}(c) = f(\rho_{ij}(c), \rho_{jk}(c)) \tag{3.13}$$

where $f(.)$ represents a rating propagation function for inference across 2 edges.

The following several sections examine 3 specific models that instantiate the above formalism. In particular, we examine the following models:

- Centrality based rating
- Preference based rating
- Bayesian inference based rating

Commercial systems such as those in Amazon or eBay are essentially simple forms of centrality based rating systems. Sociologists have studied these systems using social networks for many decades (Wasserman and Faust, 1994). In this section, we examine one particularly well constructed centrality based rating system that takes into consideration not only the ratings themselves, but also weights these ratings of the ratings on the raters themselves.

In general, centrality based rating systems are global rating systems. We argue that a personalized rating system with contextualized ratings would yield better measures of reputation or reliability. We construct two different personalized rating systems based on indirect inference. Indirect inference refers to a rating propagation mechanism to be defined below. Comparisons of these rating systems are performed in the next chapter.

3.4 Centrality-based Rating

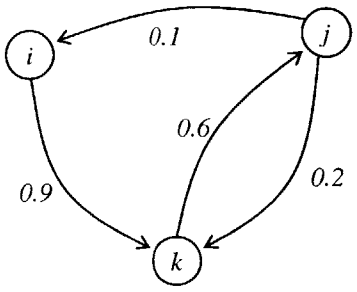


Figure 3.3 A sample social network with weighted edges representing ratings given by the source node to the destination node.

Ratings in the context of a social network can be represented by an adjacency matrix, A. A rating by member i about member j is the matrix entry a_{ij}. For the sample social network in Figure 3.3, the adjacency matrix is:

$$A = \begin{bmatrix} 1.0 & 0.0 & 0.9 \\ 0.1 & 1.0 & 0.2 \\ 0.0 & 0.6 & 1.0 \end{bmatrix} \tag{3.14}$$

We assume that every member agrees with his or her own judgment; therefore, the diagonal of the adjacency matrix is always 1.

For sociologists, network centrality is an important concept that represents how well connected an actor in a social network – his or her "reputation" (Bonacich, 1987). Simple measures of centrality involve the indegrees and outdegrees of a given member. More realistic models of prestige incorporate measures in which the centralities or prestige of nodes in a social network are recursively related to those of the nodes to which they are connected (Wasserman and Faust, 1994). Such recursive measures imply that being rated highly by a prestigious member should increase one's reputation in a social network. Being rated poorly by a reputed member should be more devastating than being so rated by a less prestigious member.

Let us define the following quantities:

$\boldsymbol{x}$: the vector of reputation measures for members in a social network

x_i : the reputation measure for member i

n : number of members in a social network

where the reputation vector $\boldsymbol{x} = [x_1 \quad x_2 \quad \ldots \quad x_n]^T$

The reputation measure x_i is a function of the reputation values of the members who have rated the member i. Note that the ith column of the adjacency matrix contains the ratings that the other members of the social network gave to member i.

The rating process by members in a social network for the ith member can be expressed as:

$$x_i' = a_{1i}x_1 + a_{2i}x_2 + \ldots + x_{ni}x_n \tag{3.15}$$

In matrix form:

$$\boldsymbol{x}' = A^T \boldsymbol{x} \tag{3.16}$$

To perform the recursive calculation referred to earlier, this equation can be solved for its steady state values (eigenvalue λ=1):

$$\left(I - \lambda A^T\right)\boldsymbol{x} = \boldsymbol{0} \tag{3.17}$$

This characteristic equation does not in general have a steady state solution since the matrix A in general does not have an eigenvalue 1.

By the Perron-Frobenius Theorem, if the adjacency matrix can be column normalized to have a unity sum, the normalized matrix will have an eigenvalue of 1 (Ross, 1995). Such normalization does not affect the relative reputation of each member in the social network. After the normalization, the largest eigenvalue of the matrix A^T is guaranteed to be 1. The corresponding eigenvector $\boldsymbol{r}$ is then the reputation of the corresponding members in the social network. This recursive calculation yields reputation that would be higher if either the ratings were higher or if the members who give those good ratings have higher reputation themselves.

Centrality based rating systems based on the above recursive procedure make use of aggregate ratings for a given context to select the most reputable members of a social network. In selecting the members with the highest reputation from the eigenvector with eigenvalue 1, the system is choosing the members whose opinions reflect that of the majority of the other members in the community. By its global nature, the same group of members of high reputation is recommended to all members in the social network.

3.5 Preference-based Rating

A preference-based rating system is a personalized rating system that takes into account the preferences of each member when selecting the reputable members in the community that he or she is most likely to approve of. Let $\rho_i(c)$ be defined as the probability that an individual i approves of an object that can be categorized within context c. The probability that i approves of another j's opinion for an object in the context c is represented by $\rho_{ij}(c)$.

Every member in a social network has a personal preference for that context. His or her preference is reflected in the way that he or she rates the other members in the community. Therefore, a subject's personal preference is inherent in the direct ratings that he or she gives. These *direct* ratings are used to estimate the ratings that he or she would give to other members in the social network. In this way, the personal preference is incorporated into the selection of reputable members of the social network.

3.5.1 Binary Pair-wise Ratings

To simplify the derivation of a rating propagation function based on preferences, we first consider binary ratings. Let member i be the subject that the preference-based system is working for. Let the personal preference of member i with respect to context c be represented by $\rho_i(c) \in \{ 0, 1\}$, where 0 indicates that member i does not approve

context c and 1 indicates that he or she does. Let member j be another member in the social network such that $i \neq j$. We consider the function that indicates how i approves of j:

$$\rho_{ij}(c) = \begin{cases} 1 & if\ \rho_i(c) = \rho_j(c) \\ 0 & otherwise \end{cases} \tag{3.18}$$

In terms of the rating propagation function introduced earlier, the rating propagation function for 2 links can be written as:

$$\begin{aligned} \rho_{ik}(c) &= f(\rho_{ij}(c), \rho_{jk}(c)) \\ &= \rho_{ij}(c)\rho_{jk}(c) + (1-\rho_{ij}(c))(1-\rho_{jk}(c)) \end{aligned} \tag{3.19}$$

To determine the propagated rating for member n that are more than 2 links away from member i, member i simply needs to apply the above calculation recursively along the path between i and n.

3.5.2 Continuous Pair-wise Ratings

Individual preferences are definitely not binary but occupy values along a spectrum. We can model individual preference using a probability value to indicate the likelihood that he or she "prefers" a given object in context c. For example, in the case of restaurant preference, one's preference $\rho_i(c)$ for spicy cuisine (where c=spicy cuisine) could vary from 0 to 1 where 0 indicates that this individual does not prefer spicy cuisine at all; to 0.5 where one is ambivalent about spicy cuisine; to 1.0 where one absolutely loves spicy cuisine. As a probability, the range of the preference function for member i is then:

$$\rho_i(c) \in [0, 1] \tag{3.20}$$

Let member $j \neq i$ be another member of the social network. Given $\rho_i(c)$, let the rating that member i would give to j, the rating that member i would give to member j with respect to context c is the probability $\rho_{ij}(c)$ that member i would approve of member j's preference for c:

$$\rho_{ij}(c) \in [0, 1] \tag{3.21}$$

Taking the spicy cuisine example from above: if member i has a preference value of $\rho_i(spicy)$ for spicy cuisine, the probability that he would approve of any randomly selected restaurant is given by $\rho_i(spicy)$. Similarly, the probability that member j would approve of any randomly selected restaurant is given by $\rho_j(spicy)$. Logically, the probability that i would approve of j's opinion on a randomly selected restaurant is the sum of the following 2 probabilities (assuming that their opinions are formed independently):

- they both approve of that restaurant: $\rho_i(spicy)\ \rho_j(spicy)$
- they both disapprove of that restaurant: $(1-\rho_i(spicy))\ (1-\rho_j(spicy))$

Consequently, the preference rating that i would give to j can be calculated as:

$$\begin{aligned}\rho_{ij}(c) &= Prob\ (i \text{ approves of } j\text{'s preference for context } c)\\ &= Prob\ (\text{both } i \text{ and } j \text{ approve of context } c) +\\ &\quad Prob\ (\text{both } i \text{ and } j \text{ disapprove of context } c)\\ &= \rho_i(c)\,\rho_j(c) + (1 - \rho_i(c))\,(1 - \rho_j(c)) \qquad (3.22)\end{aligned}$$

The actual rating that i would give to j when they actually interact is only an estimate for $\rho_{ij}(c)$ since j does not actually reveal his preference value $\rho_j(spicy)$ to i. As a result, i gives j a rating based on how much he agrees with j's approval of objects in the context c. Therefore, the social network that is formed based on direct ratings among members would yield pair-wise rating $\widehat{\rho}_{ij}(c)$, which is an estimator for the trust rating $\rho_{ij}(c)$ where $i, j \in \{ 1, \ldots, n \}$ with n members total.

3.5.3 Ratings Propagation

Consider the simple social network depicted in Figure 3.2. i has previously directly rated j; j has previously rated k. Now, i would like to estimate $\rho_{ik}(c)$ where c is any fixed context such as how one enjoys "spicy cuisine". From Equation (3.22), if i is aware of the value $\rho_k(c)$, $\rho_{ik}(c)$ can be calculated exactly. However, in real social settings, only estimates of $\widehat{\rho}_{ij}(c)$, $i, j \in \{ 1, \ldots, n \}$ are available. As a result, we have to develop a rating propagation mechanism to estimate $\rho_{ik}(c)$ from $\widehat{\rho}_{ij}(c)$ and $\widehat{\rho}_{jk}(c)$.

Section 3.3.4 has discussed the generic form of the rating propagation function. In the case of the preference framework just presented, our goal is to derive a closed form for $\rho_{ik}(c)$ given what are known to i through j. The following theorem is proved in Appendix A:

Theorem (Preference based Rating Propagation Function). With a social network setup in Figure 3.1, the rating propagation function ρ_{ik} when i and k are 2 nodes separated by a third node j is:

$$\rho_{ik} = \begin{cases} \dfrac{(2\rho_i - 1)(2\rho_i\rho_{jk} - \rho_i - \rho_{jk} + \rho_{ij}) + (1 - \rho_i)(2\rho_{ij} - 1)}{2\rho_{ij} - 1} & \text{if } \rho_{ij} \neq 0.5 \\ 0.5 & \text{if } \rho_{ij} = 0.5 \end{cases} \qquad (3.23)$$

□

The singular point of the propagation function at $\rho_{ij} = 0.5$ should yield ρ_{ik}=0.5 can be justified by interpreting 0.5 as being the least certain probability. This least certainty is warranted since there is no direct and indirect information that i can get about k.

Since true rating ρ_{ij} and ρ_{jk} are dependent on the unknowns (from i's perspective) ρ_j and ρ_k, these can be estimated from their estimators, as discussed earlier:

$$\hat{\rho}_{ik} = \begin{cases} \dfrac{(2\hat{\rho}_i - 1)(2\hat{\rho}_i\hat{\rho}_{jk} - \hat{\rho}_i - \hat{\rho}_{jk} + \hat{\rho}_{ij}) + (1 - \hat{\rho}_i)(2\hat{\rho}_{ij} - 1)}{2\hat{\rho}_{ij} - 1} & if\ \rho_{ij} \neq 0.5 \\ 0.5 & if\ \rho_{ij} = 0.5 \end{cases} \quad (3.24)$$

where $\hat{\rho}_i$ can be a simple proportional measure. For example, $\hat{\rho}_i$ can be estimated as follows: divide the total number of i's approval by the total number of ratings that i has given to that context.

When the true random variable is replaced by its estimate, equation (3.24) can yield a result that is not bounded by [0, 1]. Therefore, in order for $\hat{\rho}_{ik}$ to remain a probability value, it has to be rounded to within this range.

To estimate the rating of members that are more than 2 links away, Equation (3.24) can be recursively applied along the path linking i to those members – similar to the method used in the binary preference rating case.

3.6 Bayesian Estimate Rating

Let's assume that there are multiple encounters between member i and another member j with respect to a certain context c. During each encounter, i either approves or disapproves of member j's opinion on an object in the context c. How i perceives j in the context c depends on the history of approvals and disapprovals. For example, consider the case when c refers to "spicy cuisine". Every encounter between i and j involves their discussion on how they like a certain restaurant with reference to its spicy cuisine. i would give an approval rating to j if i considers j's opinion on a given restaurant "correct"; disapproval otherwise.

In this section, we propose a rating system based on Bayesian estimation in the approval framework just discussed.

3.6.1 Delegation of Approval: a Bayesian Inference

Let $x_{ab}(i)$ be the indicator variable for a's approval of b after the i^{th} encounter between them. If a and b have had n encounters in the past, the proportion of number of approvals of b by a can be modeled with a Beta prior distribution:

Let n = total number of encounters between a and b in the past

p = number of approvals of b by a in the past

θ = true proportion of number of approvals for b by a

$\hat{\theta}$ = estimator for θ based on all encounters between a and b

$$p(\hat{\theta}) = Beta(c_1, c_2) = \frac{\hat{\theta}^{c_1-1}(1-\hat{\theta})^{c_2-1}}{B(c_1,c_2)}$$

$$B(c_1, c_2) = \frac{\Gamma(c_1)\Gamma(c_2)}{\Gamma(c_1)+\Gamma(c_2)} \tag{3.25}$$

where c_1 and c_2 are parameters determined by prior assumptions. Assuming that each encounter's approval probability is independent of other encounters between a and b, the likelihood of having p approvals and $(n - p)$ disapprovals can be modeled as:

$$let\ D = \{\, x_{ab}(1), x_{ab}(2), \ldots x_{ab}(n) \,\}$$

$$L(D \mid \hat{\theta}) = \hat{\theta}^p (1-\hat{\theta})^{n-p} \tag{3.26}$$

where the likelihood is derived by observing that the random variable p follows a binomial distribution. Since D is the collection of all possible sets of n encounters that contains p approvals and $(n - p)$ disapprovals, its likelihood given the estimator $\hat{\theta}$ is:

$$\Pr(p, n \mid \hat{\theta}) = \binom{n}{p} \hat{\theta}^p (1-\hat{\theta})^{n-p}$$

$$L(D \mid \hat{\theta}) \propto \hat{\theta}^p (1-\hat{\theta})^{n-p} \tag{3.27}$$

Since likelihood's need not be normalized to one, the proportion is turned into an equality in Equation (3.26).

Combining the prior and the likelihood, the posterior estimate for $\hat{\theta}$ becomes:

$$\begin{aligned} p(\hat{\theta} \mid D) &= \frac{L(D \mid \hat{\theta}) p(\hat{\theta})}{\int_{\hat{\theta}} L(D \mid \hat{\theta}) p(\hat{\theta}) d\hat{\theta}} \\ &= Beta(c_1 + p, c_2 + n - p) \end{aligned} \tag{3.28}$$

The steps of derivation for Equation (3.28) are given in Appendix B. 1st order statistical properties of the posterior are summarized below for the posterior estimate of $\hat{\theta}$:

$$E\left[\hat{\theta} \mid D\right] = \frac{c_1 + p}{c_1 + c_2 + n}$$

$$\sigma_{\hat{\theta}|D}^{\ 2} = \frac{(c_1 + p)(c_2 + n - p)}{(c_1 + c_2 + n)^2 (c_1 + c_2 + n - 1)} \tag{3.29}$$

In the approval framework for reputation, reputation for b in a's mind is a's estimate of the probability that a will approve of b in the next encounter. This estimate is

based on n previous encounters between them. (Note that only a single context c_* is considered here.) This estimate can be calculated as follows:

$$p\left(x_{ab}(n+1)=1 \mid D\right)=\int p\left(x_{ab}(n+1)=1 \mid \hat{\theta}_n, D\right) p(\hat{\theta}_n \mid D)\, d\hat{\theta}_n \tag{3.30}$$

where $\hat{\theta}_n$ is the estimated approval proportion based on n previous encounters. Note that $p\left(x_{ab}(n+1)=1 \mid \hat{\theta}_n, D\right)$ is the likelihood for $x_{ab}(n+1)=1$, given the estimated parameters from n previous encounters. Substituting in the (normalized) likelihood:

$$\begin{aligned} p\left(x_{ab}(n+1)=1 \mid D\right) &= \int \frac{L\left(x_{ab}(n+1)=1 \mid \hat{\theta}_n\right)}{L\left(x_{ab}(n+1)=1 \mid \hat{\theta}_n\right)+L\left(x_{ab}(n+1)=0 \mid \hat{\theta}_n\right)} p(\hat{\theta}_n \mid D)\, d\hat{\theta}_n \\ &= \int \hat{\theta}_n p(\hat{\theta}_n \mid D)\, d\hat{\theta}_n \\ &= E\left[\hat{\theta}_n \mid D\right] \end{aligned} \tag{3.31}$$

which we have formulas for (*c.f.* Equation (3.28)). This conditional expectation is the operational definition for reputation: r_{ab}.

How is the prior belief about $\hat{\theta}$ estimated? We now consider two approaches.

3.6.2 Complete Strangers: Prior Assumptions

If individuals a and b are complete strangers, an ignorance assumption is made. When these 2 strangers first meet, their estimate for each other's reputation is uniformly distributed across the reputation's domain. *i.e.*,

$$p(\hat{\theta})=\begin{cases} 1 & 0<\hat{\theta}<1 \\ 0 & otherwise \end{cases} \tag{3.32}$$

For the Beta prior, values of c_1=1 and c_2=1 yields such a uniform distribution.

3.6.3 Known Strangers: Rating Propagation Function

If a_i and a_k have never met before but a_i knows a_j well (*i.e.*, a_i has an opinion on a_j's reputation). Also, a_j knows a_k well. a_i would like to estimate a_k's reputation based on a_i 's history of encounters with a_j, and a_j's history of encounters with a_k. This setup is depicted in Figure 3.2.

For agent a_i, agent a_k is a stranger but a_k is not completely unknown to a_i since a_i knows a_j who has an opinion about k. In a future encounter between a_i and a_k, the probability that this encounter will be rated good by a_i can be shown to be (*ibid.*):

Let $D_{ij,n}$ = all n encounters between a_i and a_j.

$D_{jk,m}$ = all m encounters between a_j and a_k.

$x_{ij}(n)$ = indicator variable for a_i's approval of a_k at encounter n

$$
\begin{aligned}
& p\left(x_{ik}(n+1)=1 \mid D_{ij}, D_{jk}\right) \\
&= p\left(x_{ij}(n+1)=1 \mid D_{ij}\right) \cdot p\left(x_{jk}(n+1)=1 \mid D_{jk}\right) + \\
&\quad \left[1-p\left(x_{ij}(n+1)=1 \mid D_{ij}\right)\right] \cdot \left[1-p\left(x_{jk}(n+1)=1 \mid D_{jk}\right)\right] \\
&= r_{ij} r_{jk} + (1-r_{ij}) \cdot (1-r_{jk})
\end{aligned}
\tag{3.33}
$$

The interpretation of this equation is that the probability that a_i would approve a_k at encounter $n + 1$ is the sum of the probabilities that both a_i and a_j agree (about the context that c is in) and that both of them disagree.

3.6.4 *Inference Propagation*

The formulation in Section 3.6.1 provides a Bayesian maximum posterior estimate of the direct neighbor reputation. The previous Section 3.6.3 has discussed a belief estimation for indirect neighbors which are one degree away from a direct neighbor. The reputation of indirect neighbors which are further away can be estimated by applying Equation (3.33) recursively along the nodes in the path that connects any two nodes in the social network. This recursion has a fixed point for a finite population community: when all members of the communities have been rated, the recursion stops.

There is a tricky issue skimmed over in the previous paragraph. What should the reputation estimate be if there are multiple paths connecting two agents? Loopy probability estimates have been found to be tricky (Pearl, 1988). Based on experimental findings to be discussed in the next chapter, the maximum probability path of the direct neighbor is chosen for inferring the reputation of indirect neighbors. Currently, we are investigating different schemes for loopy network inference (such as Murphy, *et al.*, 1999). In Chapter 5, we present another approach based on graph transformation.

3.7 Prelude to Experiments

The next chapter describes several experiments based on the rating models presented here. These experiments aim to quantitatively compare the rating algorithms against a control. One important goal of these experiments is to understand whether the commercially popular global rating systems are robust as characteristics of the underlying social network change. Two datasets are used for our experiments: one based on a set of simulated restaurant rating data, the other based on a set of real world data collected by a movie-rating web site.

CHAPTER 4

Rating Experiments

This chapter describes a series of experiments for testing the effectiveness of several rating systems based on the rating process from the previous chapter. In particular, we are interested to compare our two proposed (preference based and Bayesian estimate) personalized rating systems against existing global rating schemes (such as the eBay rating system). The first set of experiments involves a custom built simulation environment for restaurant recommendation ratings – the Restaurant Sanctioning System (RSS). The second set of experiments uses the same simulation engine on a real world dataset collected by MovieLens for movie ratings.[1] This dataset consists of 100,000 ratings by 943 users on 1682 movies of 19 different genres.

Section 4.1 describes the common experimental framework used in all experiments in this section. Section 4.2 describes the details of our restaurant rating simulation experiments. Section 4.3 describes the movie rating experiments. Section 4.4 reports the experimental results and their implications for designing rating systems and for the production of trust in online communities. Section 4.5 concludes this section and briefly discusses what has been learned from the rating model and experiments.

4.1 Experimental Framework

This section describes a system that simulates an online community of users that subscribe to a web-based service where they have to rate both resources and other users' ratings. In our simulations, the resources are restaurants and movies, and the service could be a restaurant recommendation or movie rating service that recommends restaurants or movies based on the ratings given by users of the service.

For the sake of conciseness, the description of the experimental framework below will mainly use restaurants as the resource under discussion.

4.1.1 The Simulation System

There are 4 components to the simulation system: user specifications, resource specifications, simulation engine, and analysis package. Each experiment starts with selecting the number of users and resources that are to be used. The user and resource specifications declare experimental variables and attributes (the set **B** in our rating framework) that are relevant for each experiment. The simulation engine would conduct

[1] The MovieLens web site is at: http://www.movielens.umn.edu/

the experiments according to the specifications. The results of the experiments are fed into the analysis package.

4.1.2 User Specification

A user specification (US) declares the number of users to be included in an experiment and the attributes for each user. In the case of restaurant rating, the attribute includes the home city of the user, the preference values of each user for various contexts, the number of restaurants and other members that he will rate per round of the experiment. Each user only rates restaurants that are in his or her home city. The preference values fall in the range [0, 1]. The simulation engine then generates the rating that the user would give to a restaurant for a given context based on the user's preference values as well as the restaurant's attributes. The exact rating function will be discussed in the section on the simulation engine below.

4.1.3 Resource Specification

The resource specification (RS) contains information about the size of the object set **O**, as well as the attributes of the objects themselves. In the case of restaurant rating, the attribute set include the city that a restaurant is in, the cuisine type, the quality of food in the restaurant, the ambience, *etc.* In the actual simulation, a restaurant with a low quality score for its cuisine type (the "context"), would on average receive lower rating from users for that cuisine type. Details of the rating process will be discussed below.

4.1.4 Simulation Engine

Given the user and resource specifications (and randomizer seed values), the simulation engine (SE) is a deterministic automaton that produces results that simulate how the users would rate the resources. For a given resource context (*e.g.*, Japanese restaurants in the city of San Francisco), the simulation engine makes each user rate a number of resources in that context. The set of ratings from all users then forms the social network as discussed in the rating process of Section 3.3.

How does the simulation engine determine the ratings by users on resources? Consider the following setting:

- c : context c
- x_c : a user's rating for a resource in context c
- p_c : a user's preference for a resource in context c

The simulation engine generates rating for users according to the following density for random variable $\mathbf{X}_c$ of which x_c is an instance:

$$X_c \sim \frac{1}{n} Binomial(n, p_c), \quad where\ n = 20$$
$$E(X_c) = p_c, \quad Var(X_c) = \frac{1}{n} p_c(1 - p_c) \tag{4.1}$$

The binomial density is chosen due to its ease of implementation. We could have used other density (such as Gaussian).

Since users' ratings should also be influenced by the quality of the resource for specific contexts, some modification to the above density is needed. The criteria behind the incorporation of resource quality are as follows:

- If the resource quality is very low, the mean of the rating distribution $\mathbf{X}_c$ should correspondingly be very low. For example, if the quality of the resource for the context c is $q_c = 0$, the simulation engine can choose $E(\mathbf{X}_c) = 0$.
- Conversely, when the resource quality is high, we expect $E(\mathbf{X}_c)$ to be high. For example, if the resource quality for the context c is $q_c = 1$, the simulation engine can choose $E(\mathbf{X}_c) = 1$.

To satisfy these guidelines, we adopt the following variation on the mean of X_c.

$$X_c \sim \frac{1}{n} Binomial(n, \mu_c), \quad where\ n = 20$$
$$E(X_c) = \mu_c = p_c + 2(q_c - \frac{1}{2})(q_c - p_c) \tag{4.2}$$

Note that when $q_c = 0.5$, the mean value of $\mathbf{X}_c$ is p_c. Such a rating distribution satisfies the two criteria listed above.

For every user, the simulation engine draws a rating from $\mathbf{X}_c$ for every resource that user is to rate. The set of all user ratings are the inputs to the analysis package.

4.1.5 Analysis Package

Based on the attributes of the users, the resources, the ratings of resources by users, the analysis package constructs a social network for every context of interest. In other words, a social network is formed among individuals who happen to rate resources for the context of interest (such as "spicy cuisine"). The social network is the basis for analyzing the effectiveness of various rating and rating propagation schemes.

4.1.6 Error Measure for Analysis

The error measure will be the 'ranking error measure', described as follows. For every user, the simulation run would use the propagation mechanism of the rating system to generate the ratings for his indirect neighbors. Using these ratings and the direct ratings given to his first-degree friends, a list of neighboring users is generated. The inferred reputation ratings for all indirect neighbors are used to rank these neighbors. Next, we compare the ranked list of all users each rating system generates to the correct ranked list. The correct ranked list is generated by calculating the direct rating that the subject user would have given to all users in the network, and then ranking all these users using these direct ratings. The ranking error measure calculates the difference in the ordering of users in the two lists. It does that by first finding the sum of the absolute differences of each user's ranking in the two lists. This sum is then normalized by dividing it by the maximum possible error for the number of users. Note that the maximum error occurs when the inferred ranking is exactly opposite to that of the actual one:

$$\text{Max(ranking error measure)} = \begin{cases} n^2 - 1/2, & n \in Odd \\ n^2/2, & n \in Even \end{cases} \tag{4.3}$$

where n = number of users in the community. The ranking error measure is therefore a fraction given by the normalized value.

The ranking error measure is a modified measure based on the Wilcoxon sign test from statistics. Estimators based on the Wilcoxon sign test has a range of all 0 and positive integers. The ranking error measure has been normalized by (4.3) to have a range of [0, 1].

4.2 Restaurant Rating Simulation

The simulation environment is used to simulate restaurant ratings by users. The preference values of the users for given contexts are drawn from the range [0, 1]. Every user randomly selects a set of restaurants, and gives restaurant ratings with respect to specific contexts. When user a_i decides to rate user a_j for the context of ambience, user a_i will first ask for the set of restaurants that user a_j has rated with respect to the context of ambience. He then determines the set, D_{ij} of restaurants that both of them have rated. Using D_{ij} and a_j's ratings for restaurants in D_{ij}, a_i then determines rating $r_{ij} \in R_{ij}$ for a_j. The rating r_{ij} that a user a_i gives to another user a_j confers the reputation rating that a_j receives from a_i. r_{ij} should reflect how much a_i approves of a_j's ratings of restaurants in the set D_{ij}. In personalized rating systems, we expect r_{ij} to depend on the number of restaurants that a_i and a_j share in common as well as their opinions of these shared restaurants.

4.2.1 Level of Approval

How is the level of approval determined given a set of restaurant ratings by a_j on the set of restaurants D_{ij}? Since the ratings range from [0, 1], there is no straightforward way to determine such agreement. We present two heuristic algorithms here.

4.2.1.1 Threshold Algorithm

The Threshold Algorithm essentially turns real value ratings into binary ratings. It makes an arbitrary boundary (midpoint in the range) that separates one binary value from another. The complete threshold algorithm is shown below:

Threshold Algorithm

Given: ratings $\rho_i(r_k) \in [0, 1]$ and $\rho_j(r_k) \in [0, 1]$ that agent i and j have respectively given to restaurant $r_k \in D_{ij}$

Let $p = 0$ and n = size of D_{ij}.

For r_k in D_{ij}

if $((\rho_i(r_k) \geq 0.5 \wedge \rho_j(r_k) \geq 0.5) \vee (\rho_i(r_k) < 0.5 \wedge \rho_j(r_k) < 0.5))$

$p = p + 1$

$r_{ij} = p/n$

For example, if user A and user B both give a restaurant good rating ($\rho(r_k) \geq 0.5$), the Threshold Algorithm considers user A approves of B's rating. Consequently, p in the algorithm will be incremented by 1.

4.2.1.2 Agreement Likelihood Algorithm

We can also assess the level of approval that an agent i has for another j's ratings as the probability that the i would approve of j's ratings with respect to a given context. If $\rho_i(r_k)$ represents the probability that i likes the restaurant r_k, and $\rho_j(r_k)$ represents the probability that j likes the restaurant r_k. The probability that they both like r_k is given by $\rho_i(r_k)\ \rho_j(r_k)$ assuming that they each arrive at their assessment independently. Similarly, the probability that they both dislike r_k is given by $(1 - \rho_i(r_k))\ (\ 1 - \rho_j(r_k))$. Therefore, the probability that i would approve of j is the sum of these two products. Finally, the rating that i has for j would be the proportion of restaurants in D_{ij} that i would approve of j's ratings.

Agreement Likelihood Algorithm

Given: ratings $\rho_i(r_k) \in [0, 1]$ and $\rho_j(r_k) \in [0, 1]$ that agent i and j have respectively given to restaurant $r_k \in D_{ij}$

Let $p = 0$ and n = size of D_{ij}.

For r_k in D_{ij}

$$p = p + \rho_i(r_k)\ \rho_j(r_k) + (1 - \rho_i(r_k))\ (1 - \rho_j(r_k))$$

$r_{ij} = p/n$

4.2.2 Rating Propagation Algorithms

Either of the two algorithms above can provide user-to-user direct ratings between any pair of users. In most social scenarios, it is unlikely that every encounters would involve interactions between previously rated agents. Therefore, some form of rating propagation is necessary to spread ratings across a social network of agents. In the example in Figure 3.2 where edges refer to the existence of direct ratings, for one to infer how i might rate k, one would have to estimate r_{ij} based on r_{ij} and r_{jk}.

The user-to-user direct ratings are used to construct a social network for the context c being considered. This social network is used for propagation of ratings. 4 rating propagation schemes are experimentally tested. In the description of these schemes below, note that first degree direct neighbors of a user i refer to all those that i has directly rated; second degree indirect neighbors are those that are two rating links away. Indirect neighbors refers to all neighbors that are at least second degree and above. These 4 schemes are:

1. **Control rating**: As a control, indirect neighbors are randomly ranked. Note that when the number of first degree direct neighbors equals the total number of users – 1, ranking error (*c.f.*, Section 4.1.6) should be 0 since there are no second degree indirect neighbors in this case.
2. **Global rating**: The most reputed individual is computed using the centrality measure based on eigenvector method discussed in Section 3.4. To simulate a reliance on a "global" reputation measure, every user uses the ranking of this reputed individual as

the ranking of their indirect neighbors. This global rating scheme is similar in nature to those used in many commercial communities such as eBay.

3. **Preference-based rating**: The Preference based Rating Propagation Function from Section 3.5 is used recursively across the nodes in the path between all rating agents to infer all indirect ratings:

$$\hat{\rho}_{ik} = \begin{cases} \dfrac{(2\hat{\rho}_i - 1)(2\hat{\rho}_i\hat{\rho}_{jk} - \hat{\rho}_i - \hat{\rho}_{jk} + \hat{\rho}_{ij}) + (1-\hat{\rho}_i)(2\hat{\rho}_{ij} - 1)}{2\hat{\rho}_{ij} - 1} & \text{if } \rho_{ij} \neq 0.5 \\ 0.5 & \text{if } \rho_{ij} = 0.5 \end{cases} \quad (4.4)$$

where $\hat{\rho}_i$ and $\hat{\rho}_i$ are agent i and j's estimated preference values for a given context. $\hat{\rho}_{ij}$ and $\hat{\rho}_{jk}$ are the direct ratings between agents i and j, and j and k respectively. For individual preference estimation, $\hat{\rho}_i$ is calculated as the average of all the restaurant ratings that user i has given to restaurants in a given context.

4. **Bayesian estimate rating**: Each user infers the reputation of second degree indirect neighbors by the Bayesian method discussed in Chapter 3. The rating propagation scheme here is given by:

$$\rho_{ik}(c) = \rho_{ij}\,\rho_{jk} + (1 - \rho_{ij})\,(1 - \rho_{jk}) \quad (4.5)$$

4.2.3 Multiple Paths and Loops

A general social network is bound to have loops in the underlying undirected graph. For the personalized rating algorithms being studied, each path connecting 2 nodes is likely to produce different ratings. Consider the following undirected graph underlying a social network:

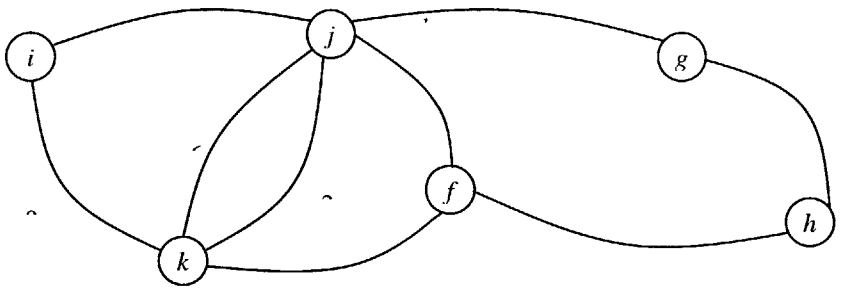


Figure 4.1. Shown here is the underlying undirected graph of a social network.

Multiple paths exist between node i and node h. In general, dealing with "loopy networks" involve one of 3 techniques (Pearl, 1988): (1) clustering : collapsing multiple nodes into a compound node, (2) cutset conditioning : changing the connectivity of a network and render it singly connected by instantiating a selected group of variables, and (3) stochastic methods : use sampling techniques and computational cycles to average over many samples of indirect inference ignoring the existence of multiple paths. Murphy, *et al.* (1999) have demonstrated that by ignoring multiple paths and dependency among variables, stochastic methods can approximate the true solution surprisingly well.

We simulated 3 different strategies for indirect inference:

- **Average strategy**: if there are m different paths between node i and j, this strategy determines i's inference for j as the simple average across the m paths.
- **Weighted strategy**: for the m different inferences, each is weighed by i's rating for the intervening node k, ρ_{ik}.
- **Maximum strategy**: for the m different paths between node i and j, pick the indirect inference from the intervening node k such that $\rho_{ik} \geq \rho_{ik'}$ for all k, $k' \in \{$ i's direct neighbors lying in the m paths connecting i and j $\}$ and $k \neq k'$.

In all 3 strategies, if the number of paths between two nodes becomes 1, the calculation degenerates to the 1 indirect inference between i and j, as expected.

4.3 Movie Rating Experiments

A set of experiments similar to those described in the restaurant case in the previous section are performed on the MovieLens data set [2]. This dataset consists of 100,000 ratings by 943 users on 1682 movies of 19 different genres. The data were collected through the MovieLens web site (movielens.umn.edu) during the seven-month period from September 19th, 1997 through April 22nd, 1998.

Each experiment assigns each user a set of direct neighbors. The task is to estimate the ranking of indirect neighbors in terms of reputation to the user concerned – similar to the restaurant simulation in the last section. The ranking error measure described earlier is also used. Referring to the rating model in Section 3.3.1, the set of agents **A** that we have is the set of 943 users in the GroupLens dataset. The set of objects **O** is the set of movies being rated by these users. The set of ratings by members of **A** on members of **O** are normalized so that each rating value falls in the interval [0, 1].

[2] MovieLens. http://movielens.umn.edu/

4.4 Experimental Results

4 sets of experiments are performed. The first 3 sets aim to evaluate the effectiveness of the rating propagation algorithms discussed in Section 4.2.2. We are particularly interested in comparing how well different personalized rating systems do compare to the commercially prevalent global rating system. Our hypothesis is that tastes can vary substantially from individual to individual in the real world; such taste differences justify the need to personalized ratings for every user. We expect personalized rating systems to be more effective (lower ranking error measure) than a global one. Through the experiments in this section, we would like to test this hypothesis under varying social network conditions.

The last set of experiments aims to evaluate the 3 strategies for dealing with multiple paths connecting two indirect neighbors.

4.4.1 Rating Propagation: Network Density Variation

The first set of experiments deal with evaluating the 4 rating propagation systems in ranking neighboring users as the density of the network varies. Density refers to the number of other users that each user rates – each rating adds an edge to the social network. The simulation engine creates an environment with 300 restaurants and each user is given 100 restaurants to rate – controlling for the statistical significance of restaurant preference values for each user.

We first use the Threshold Algorithm from 4.2.1.1 to determine direct user to user rating. Figure 4.2 shows the ranking error measure (*c.f.*, Section 4.1.6) for the 4 rating propagation algorithms as a function of neighborhood size. Note that as the neighborhood size increases, the network density increases; the converse is true when the neighborhood size decreases. Let k be the neighborhood size. For each k, 10 simulation runs are executed for each rating propagation algorithm and the average error measure for the algorithm is calculated and plotted as shown. The total number of users is 40 and k is varied from 1 to 39. Also shown in Figure 4.2 is the "Control" rating propagation algorithm. Ratings under "Control" are calculated for each user i by randomly inserting i's second-degree friends into a ranked list of i's first-degree friends. This control acts as the baseline against which we can compare the other rating propagation schemes. As long as a propagation scheme can rank the second-degree neighbors better than the "Control", we expect the scheme to be better than chance.

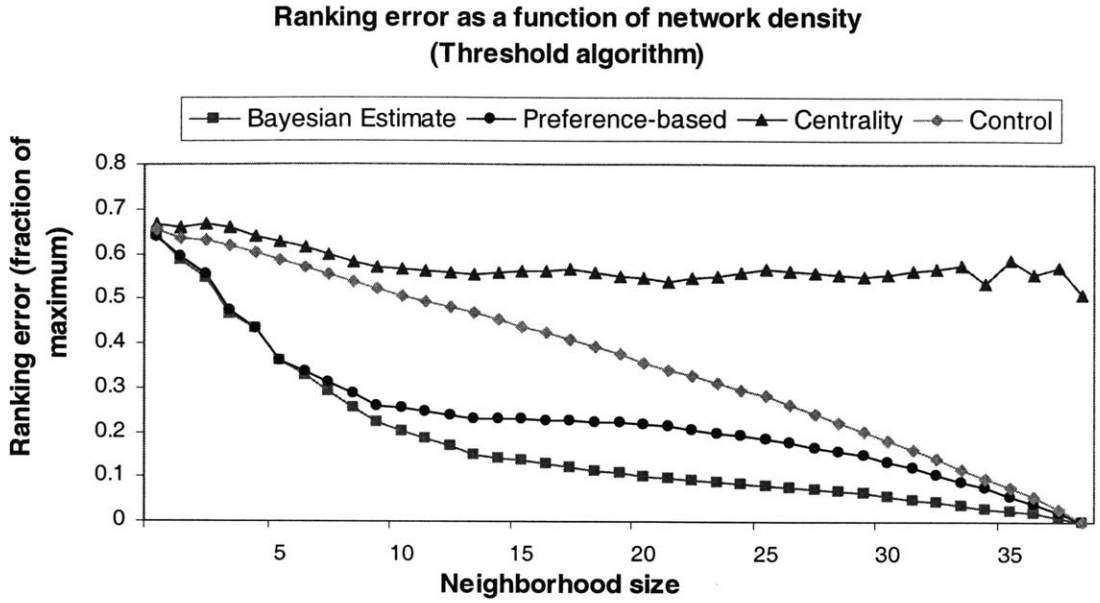


Figure 4.2. Shown is the restaurant simulation result with ranking error as a function of neighborhood size for the 4 rating propagation algorithms studied. The direct rating algorithm used is the threshold algorithm. Network density refers to number of edges over number of user nodes in a social network

Note that when the neighborhood size reaches (*n-1*) where n = size of a social network, we expect the personalized rating propagation schemes to have 2^{nd} degree neighbor ranking error measure = 0, as shown in Figure 4.2. The plots in Figure 4.2 indicate that the Bayesian estimate propagation scheme seems to achieve the least ranking error measure among the schemes studied.

When the number of direct neighbors is 0, clearly, no algorithm can perform better than random guessing. However, as each user gets to know the true reputation of additional direct neighbors for a specific context of rating, more information leads to decreased ranking error. The Centrality-based scheme consistently performs worst among all schemes examined, including the Control. What is interesting is that by trusting the most reputed individual for ranking users, each user does worse than by relying on his or her own judgment on the direct neighbors while random guessing on indirect neighbors. This should be no surprise since our reputation measure concerns degree of approvals between two agents for a specific context. This problem is especially noticeable for dense networks where information about one's neighbors is easily noticeable.

In the next experiment, we use the Agreement Likelihood Algorithm from Section 4.2.1.2 to determine direct user to user rating. Figure 4.3 plots the ranking error measure (*c.f.*, Section 4.1.6) for the 4 rating propagation algorithms as a function of neighborhood size. Note that as the neighborhood size increases, the network density increases; the converse is true when the neighborhood size decreases. Let k be the neighborhood size. For each k, 10 simulation runs are executed for each rating propagation algorithm and the average error measure for the algorithm is calculated and plotted as shown. The total number of users is 40 and k is varied from 1 to 39. Much similarity exists between

Figure 4.2 and Figure 4.3. The centrality-based scheme continues to perform poorly compare to the other propagation schemes. As the network density increases, both the Bayesian and the Preference based propagation schemes perform significantly better than the control. But in this Agreement Likelihood direct rating calculation, a Preference-based propagation scheme outperforms the Bayesian scheme. The relative performance of the preference-based versus the Bayesian schemes is discussed in Section 4.5.

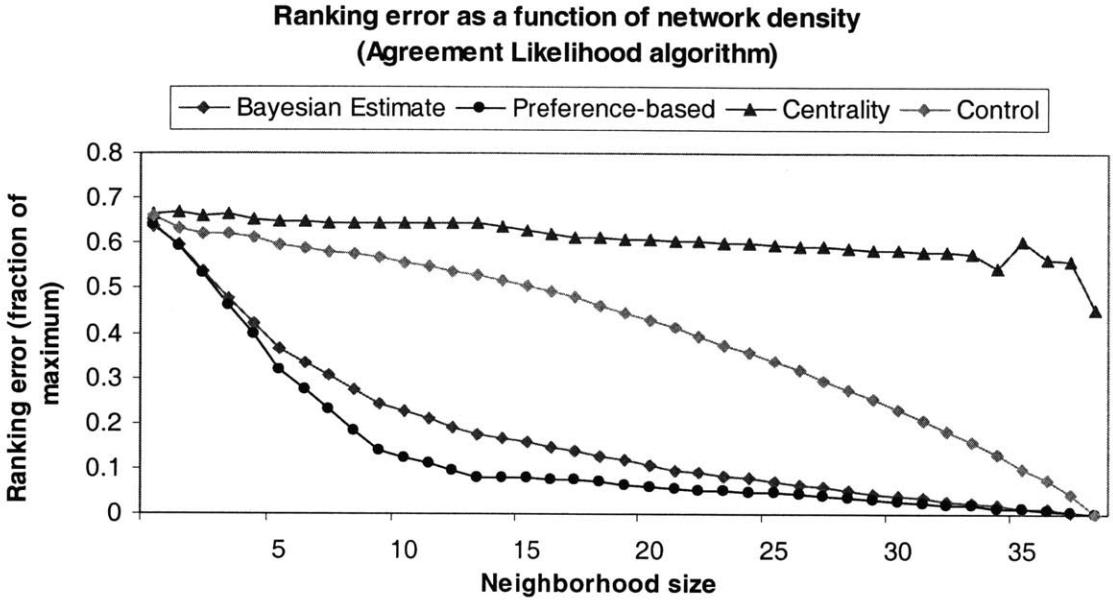


Figure 4.3 Shown is the restaurant simulation result with ranking error as a function of neighborhood size for the 4 rating propagation algorithms studied. The direct rating algorithm used is the Agreement Likelihood algorithm. Network density refers to number of edges over number of user nodes in a social network

Using the MovieLens movie rating dataset, each of the rating propagation algorithms is evaluated against these real world data. In Figure 4.4 and Figure 4.5, the x-axis indicates the number of direct neighbors (ndr) each user has. As each user gets to know the true reputation of additional direct neighbors for a specific context of rating, more information again leads to decreased ranking error. Except for small neighborhood size, the global reputation measure again leads to the most error while the Bayesian contextualized reputation algorithm leads to the least error of the three. The shapes of the ranking error curve are to first approximation the same as those derived from the restaurant simulations in Figure 4.3. Note that two different contexts are considered in the experiments shown in Figure 4.4 and Figure 4.5. In Figure 4.4, the context is for all movies in the 'Drama' genre. In Figure 4.5, the context is for all movies in the 'Romance' genre.

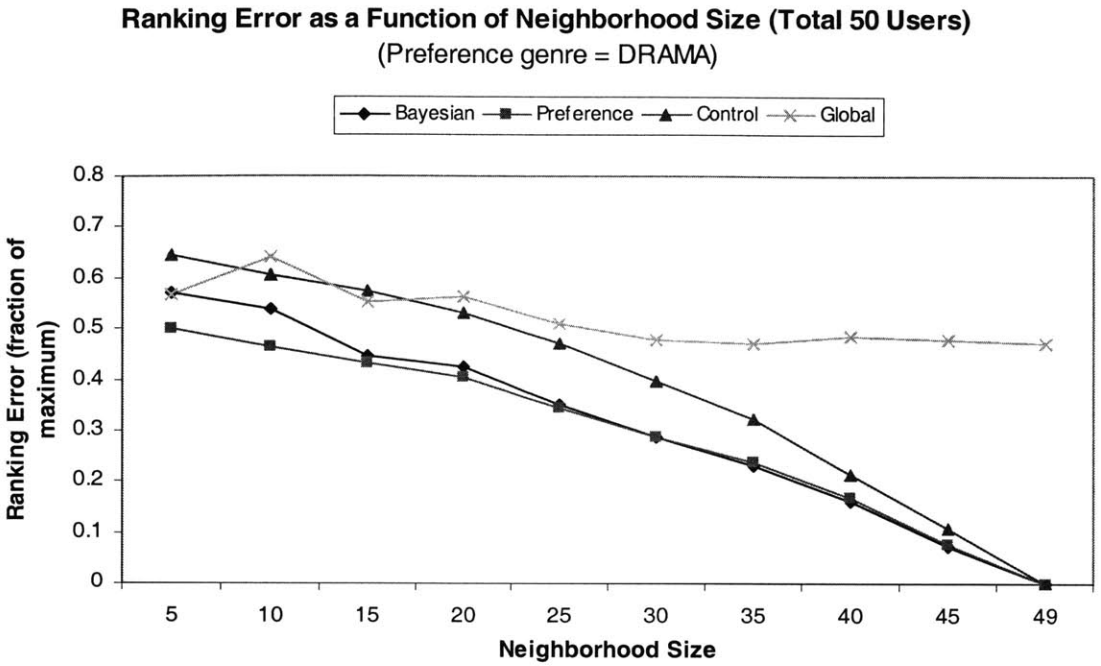


Figure 4.4. Shown are the MovieLens experimental results for the context 'Genre Drama' with ranking error as a function of neighborhood size for the 4 rating propagation algorithms studied. The direct rating algorithm used is the Agreement Likelihood algorithm.

Qualitatively, the two experiments in Figure 4.4 and Figure 4.5 have very similar results. In both cases, the Centrality-based scheme consistently performs worst among all schemes examined, including the Control. As the network density increases, both the Bayesian and the Preference based propagation schemes perform significantly better than the control. Just as the restaurant simulation results using Agreement Likelihood direct rating calculation in Figure 4.3, a Preference-based propagation scheme outperforms the Bayesian scheme.

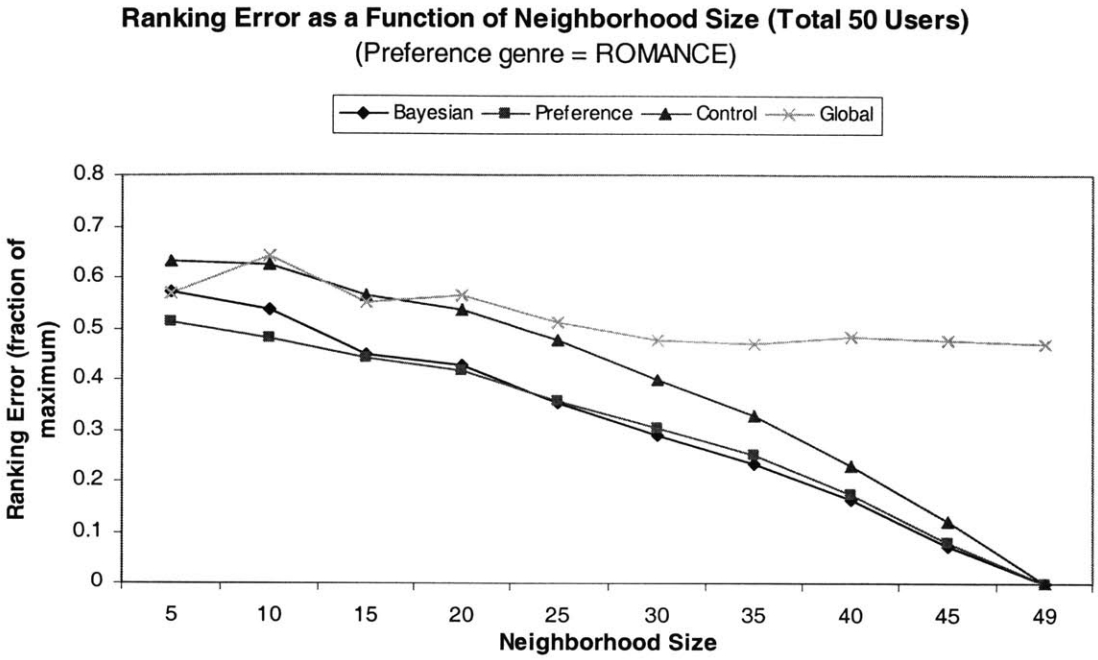


Figure 4.5. Shown are the MovieLens experimental results for the context 'Genre Romance' with ranking error as a function of neighborhood size for the 4 rating propagation algorithms studied. The direct rating algorithm used is the Agreement Likelihood algorithm.

4.4.2 Rating Propagation: Network Size Variation

A set of simulations is performed with varying number of users in a social network. The main goal of this set of experiments is to test the scalability of the 4 rating propagation schemes. In other words, we would like to know if the performance observed in the first set of experiments in Section 4.4.1 can be maintained as the number of nodes in a social network increases.

A total of 300 restaurants are defined for these experiments. Each user specification mandates each user to rate 100 restaurants (randomly chosen) so as to get fairly dense social networks. Network size is varied from 20 to 130. For each network size, 10 simulations runs are executed for each of the 4 rating propagation schemes. The line denoted "Control", as in the previous set of experiments, is the ranking error generated by the control rating propagation scheme.

Plotted in Figure 4.6 and Figure 4.7 are the ranking error curves when half of nodes in a social network are direct neighbors. Figure 4.6 shows the results using the Threshold direct neighbor rating algorithm and Figure 4.7 shows the results using the Agreement Likelihood direct neighbor rating algorithm. The results in these two figures do not seem to vary greatly as the network size varies from 20 to 130.

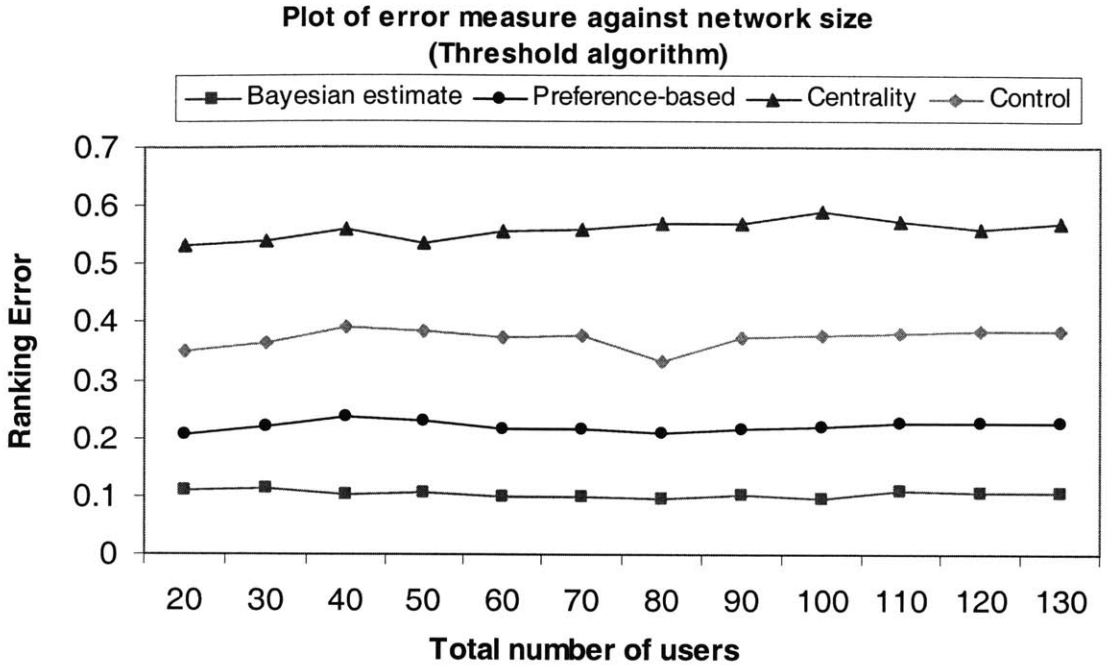


Figure 4.6. Ranking error curves for the simulated restaurant rating experiments when half of nodes in a social network are direct neighbors. The direct rating algorithm used is the Threshold algorithm.

The consistency of the ranking error measures suggests that the rating propagation schemes tested here are scalable in the range of 20 to 130 nodes in the social network. The Bayesian estimate and Preference-based rating propagation schemes perform several times better than the control system across node size in this range.

An important difference highlighted by Figure 4.6 and Figure 4.7 is that depending on the direct neighbor rating algorithm used, either the Bayesian estimate rating propagation scheme or the Preference-based rating propagation scheme achieves the least ranking error measures consistently across the network size varying from 20 to 130 nodes.

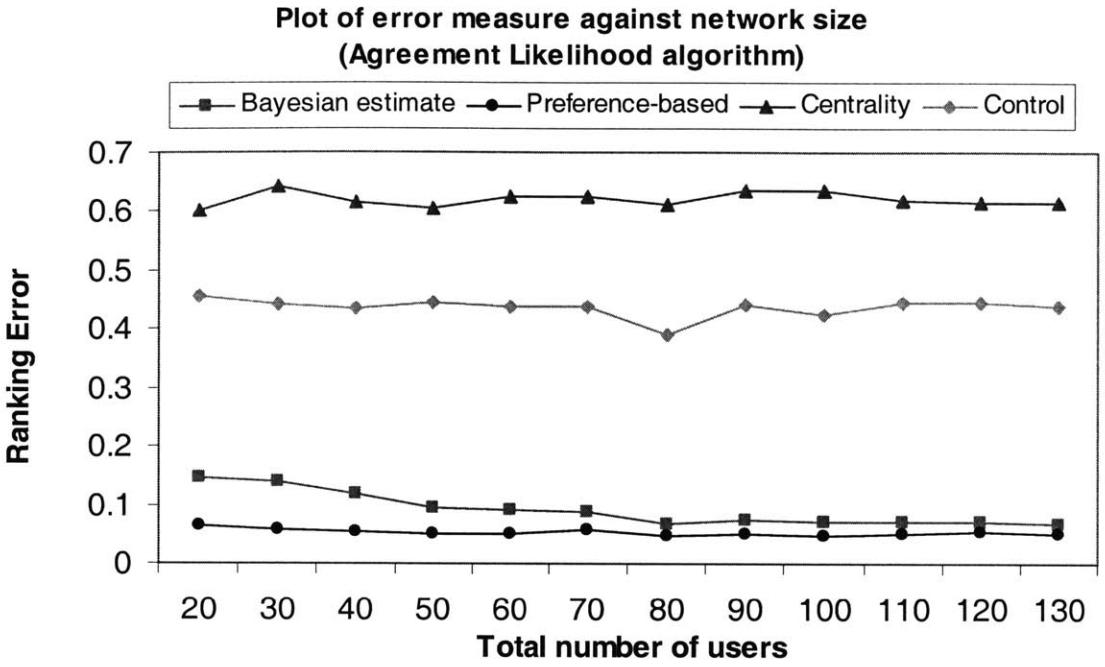


Figure 4.7. Ranking error curves for the simulated restaurant rating experiments when half of nodes in a social network are direct neighbors. The direct rating algorithm used is the Agreement Likelihood algorithm.

Performing the same scalability experiments on the MovieLens dataset, the results are shown in Figure 4.8. Just as in the case of the simulation experiments displayed in Figure 4.6 and Figure 4.7, the network size is varied while the number of direct neighbors is kept at 50%. The performance of the Preference-based and the Bayesian estimate rating propagation schemes seems to be consistent with what are observed in the restaurant simulations just discussed in Figure 4.6 and Figure 4.7: these personalized rating schemes significantly outperform both the control and Centrality-based global rating propagation system.

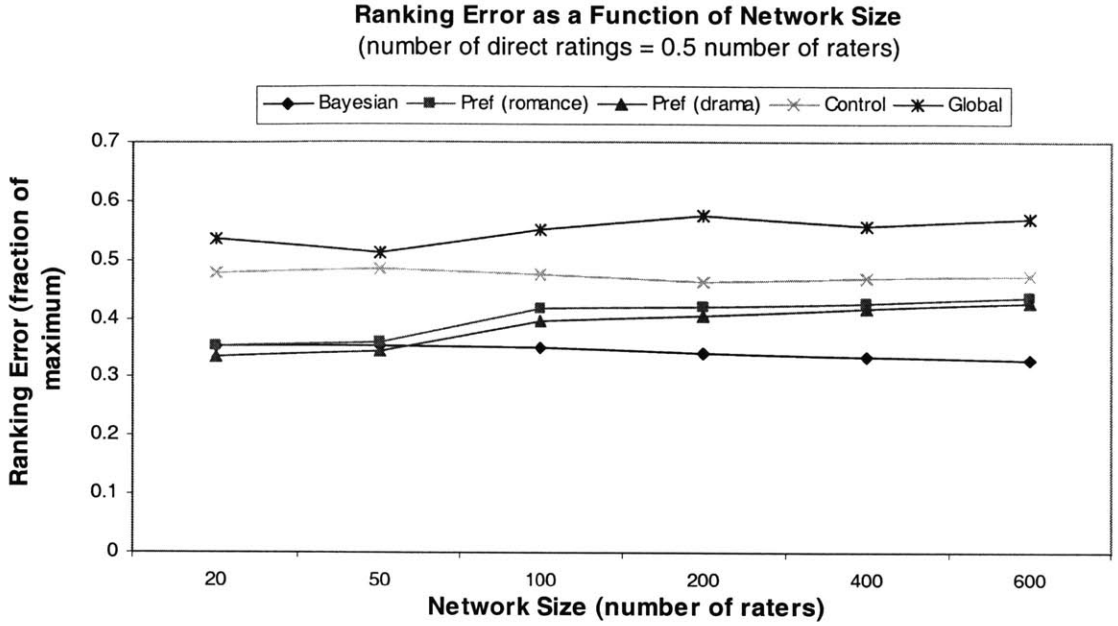


Figure 4.8. Ranking error curves for the MovieLens dataset experiments when half of nodes in a social network are direct neighbors. The direct rating algorithm used is the Agreement Likelihood algorithm.

4.4.3 Rating Propagation: Sampling Size Variation

For all experiments studied so far, the number of direct neighbors for every node in the social network is assumed to be uniform. We would like to know how reliable are the observations made so far when this assumption is removed. In other words, we are varying the accuracy of the direct neighbor ratings. The accuracy of the direct neighbor ratings depends on the number of restaurants rated by each user. As the number of restaurants rated by each user is decreased, the number of restaurants commonly rated by any two randomly chosen users decreases correspondingly. In turn, the accuracy of the rating that a user would give to another should also decrease. This set of experiments is crucial in understanding scenarios such as when a social network is being formed.

For the results shown in Figure 4.9 and Figure 4.10, 40 users are defined, with the number of direct neighbor ratings 20 per user. Each user rates between 5% to 50% of a total of 300 restaurants defined for the experiments. As in the previous cases, the line denoting "Control" is generated by a random process of inserting second degree neighbors into ranked first degree neighbors.

Figure 4.9 shows the results using the Threshold direct neighbor rating algorithm as we change the proportion of restaurants rated by each user. The ranking error for the Centralized-based global rating scheme remains constant through changes in the number of restaurants rated by users. The ranking errors from the other rating propagation schemes start off at a high value and drop as the number of rated restaurants per user increases. However, the marginal increase in the performance of the rating propagation schemes decreases with the increase in the accuracy of the direct neighbor ratings. Overall, the Bayesian estimate rating propagation scheme has the best performance in terms of the minimum amount of ranking error, followed by the Preference-based rating

propagation scheme. Consistently, the Centrality-based rating scheme has the worst performance compare to the other schemes studied, including the control.

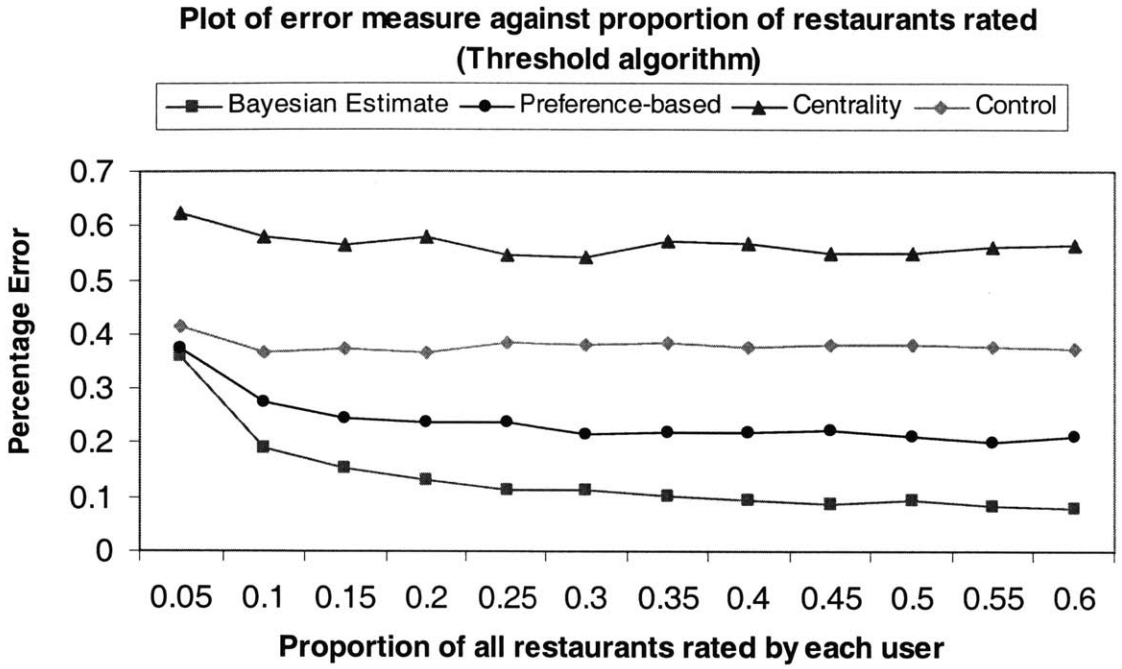


Figure 4.9. Ranking error measure as a function of the proportion of all restaurants rated. The direct neighbor rating algorithm used is the Threshold Algorithm.

Figure 4.10 shows the results using the Agreement Likelihood direct neighbor rating algorithm as we change the proportion of restaurants rated by each user. The ranking error generated by the Centrality-based rating scheme is again the worst among those examined. All the personalized rating propagation schemes start off with relatively high ranking errors. In contrast to results based on the Threshold direct neighbor rating algorithm, the Preference-based rating propagation scheme does slightly better than the Bayesian estimate propagation scheme in terms of smaller ranking errors across all proportions of restaurants rated by each user. Again, the relative performance of the preference-based versus the Bayesian schemes is discussed in Section 4.5.

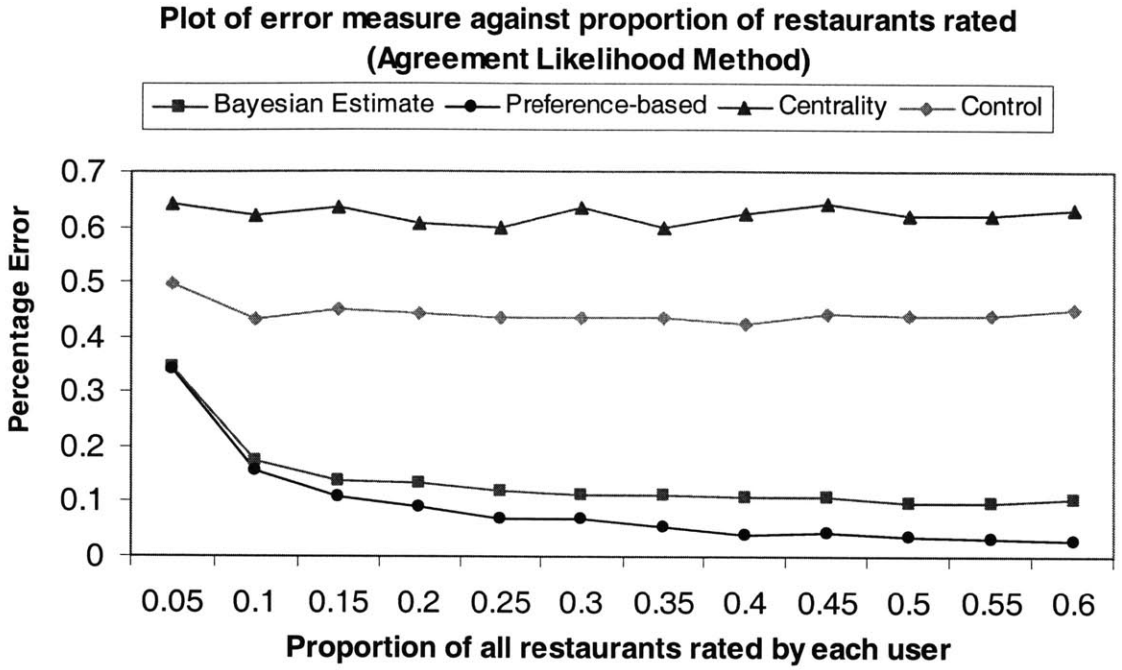


Figure 4.10. Ranking error measure as a function of the proportion of all restaurants rated. The direct neighbor rating algorithm used is the Agreement Likelihood Algorithm.

With the MovieLens dataset, the next experiment varies the network size but keeps the number of direct neighbors fixed at 25 individuals. As the number of individuals in a community increases, one expects the error in reputation ranking of indirect neighbors to increase, as is the case shown by the MovieLens results in Figure 4.11. When the number of individuals is below 400, Preference-based rating propagation achieves the least ranking error. However, as the number of individuals increases above that, the Centrality-based global reputation inference seems to have an advantage. This small observation has implications on the relationship between the importance of opinion leaders and network size and warrants future research.

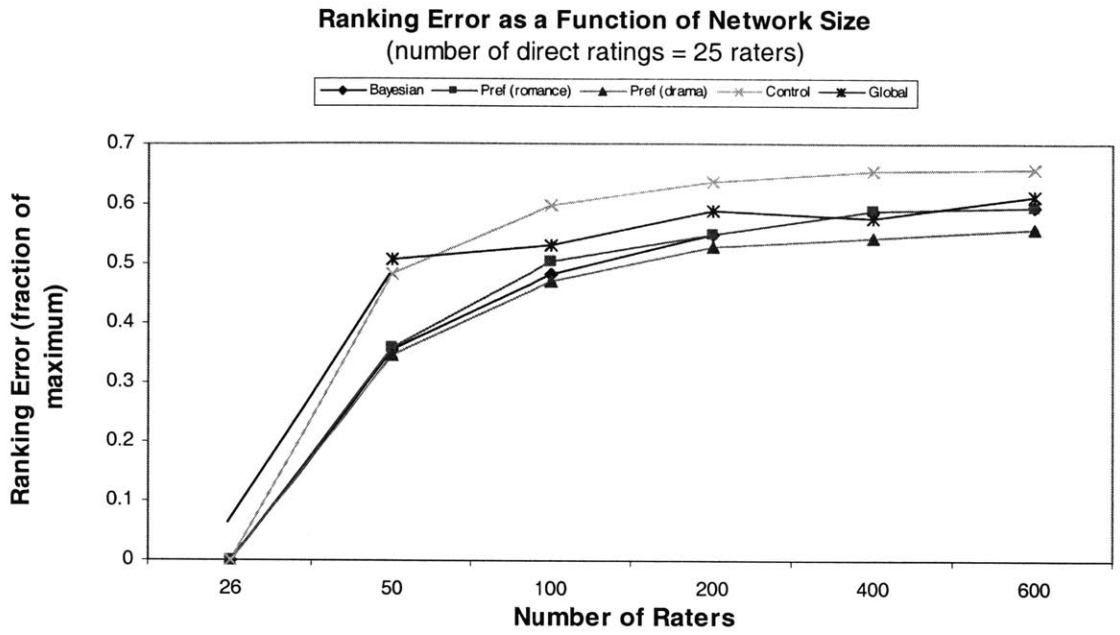


Figure 4.11. The left figure shows movie raters reputation ranking error as a function of network sizes and fixed number of direct neighbors (25)

4.4.4 Multiple Paths

We would like to examine different strategies for combining evidence from multiple paths connecting two second-degree indirect neighbors. In calculating the rating that one user would give to another that is two or more links away, we have to deal with cases where more than one path connects the pairs of users. In general, we expect different paths to produce different ratings using the rating propagation schemes discussed so far. Therefore, we must devise strategies to combine the ratings derived from these multiple paths in a meaningful manner.

In this set of experiments, we use the 3 strategies discussed in Section 4.2.3. For the following discussions, consider the following stylized version of multiple 2^{nd} degree neighbors as shown in Figure 4.12.

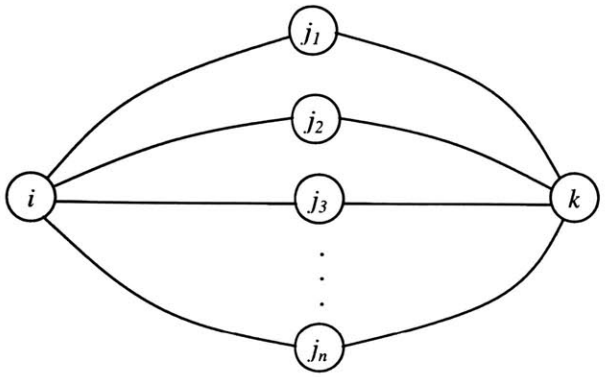


Figure 4.12. Stylized social network to illustrate multiple paths connecting two 2^{nd} degree neighbors.

For the averaging strategy, the rating r_{ik} that agent i has for k in the social network as shown in Figure 4.12 is calculated as:

$$r_{ik,averaging} = \frac{1}{n}\sum_{h=1}^{n} r_{ik}(via\ node\ j_h) \tag{4.6}$$

For the weighted averaging scheme, r_{ik} is calculated as follows:

$$r_{ik,weighted} = \sum_{h=1}^{n} r_{ij_h} r_{ik}(via\ node\ j_h) / \sum_{h=1}^{n} r_{ij_h} \tag{4.7}$$

For the maximum rating scheme, r_{ik} is calculated as follows:

$$r_{ik,maximum} = \max_{h}\{r_{ik}(via\ node\ j_h)\} \tag{4.8}$$

In the previous simulations, a 2nd degree neighbor's rating is determined using the maximum rating scheme. The following set of experiments attempt to compare the results across all 3 aggregation schemes.

4.4.4.1 Multiple Paths in Movie Ratings

We use the MovieLens dataset under the same experimental setup as in the experiments so far described. Plotted in Figure 4.13 and Figure 4.14 are the simulation results for ranked error measures for rating second degree neighbors.

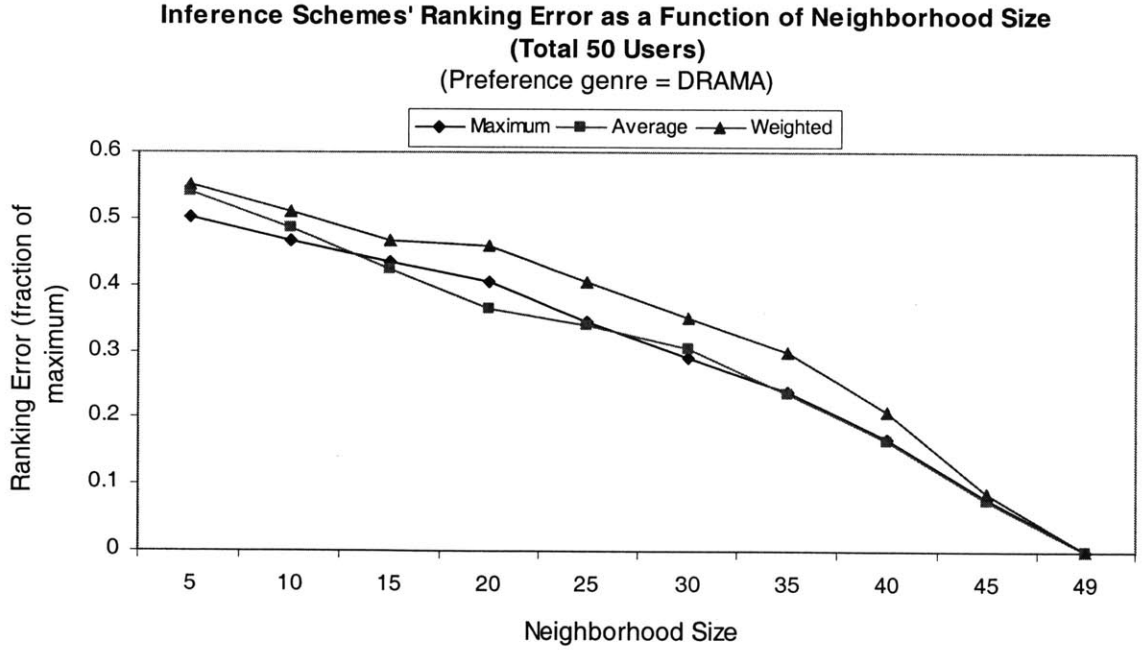


Figure 4.13. Shown is ranked error measure for the 3 schemes for combining multiple paths' inference results against direct neighborhood size. The preference-based rating propagation scheme is used for inferring indirect neighbor pair's ratings; the direct neighbor rating algorithm used is the Agreement Likelihood Algorithm. The "context" used is for ratings in the MovieLens dataset about the "genre drama".

Note that in both set of experiments using the preference-based and the Bayesian rating propagation schemes, the weighted strategy for combining multiple paths' inferences under-performs the maximum- and averaging strategies. In the case of preference-based rating propagation schemes, the maximum strategy further differentiates itself by outperforming the other two strategies in the range of neighborhood size from [5, 48].

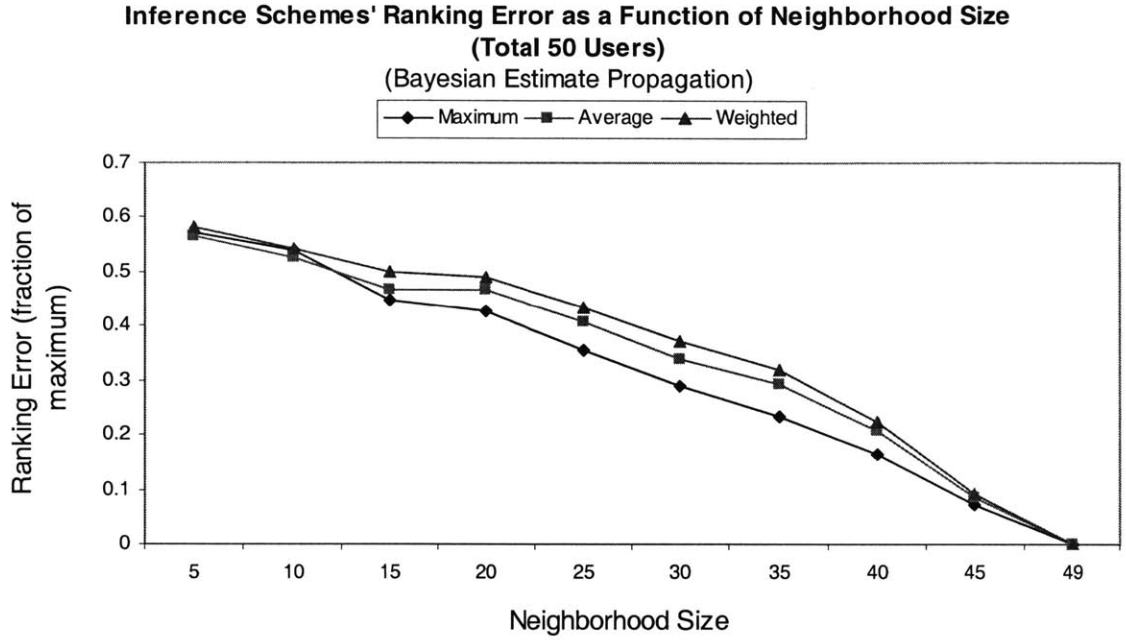


Figure 4.14. Shown is ranked error measure for the 3 schemes for combining multiple paths' inference results against direct neighborhood size. The Bayesian estimate rating propagation scheme is used for inferring indirect neighbor pair's ratings; the direct neighbor rating algorithm used is the Agreement Likelihood Algorithm.

To further study the 3 aggregation strategies for multiple paths, we set up the restaurant rating simulations as before. Every user randomly selects a specified set of restaurants for ratings. In each simulation run, the ratings that users give for the restaurants in a given context are used to construct a social network for inferring user-to-user ratings. Direct ratings use either the Threshold algorithm or the Agreement Likelihood algorithm (*c.f.*, Section 4.2.1). The 3 multiple paths aggregation strategies are simulated for both the Bayesian estimate and Preference-based rating propagation schemes.

For this set of restaurant simulations, we use a slightly different error measure ("difference error") to highlight the purpose of comparing the 3 aggregation schemes. For every agent i in the social network, the simulation run would use a rating propagation scheme to generate the estimated ratings for all i's second degree neighbors. The simulation then calculates the direct ratings that i would give to all his second-degree neighbors using either the Threshold algorithm or the Agreement Likelihood algorithm. The error measure (difference error) we use to compare across the different aggregation strategies is defined to be the average difference between each pair of direct rating and

estimated rating. Therefore, we expect this difference error measure to be in the range [0, 1]. A difference error of 1 indicates that the estimated rating completely differs from the direct ratings whereas a difference error of 0 means that they are identical.

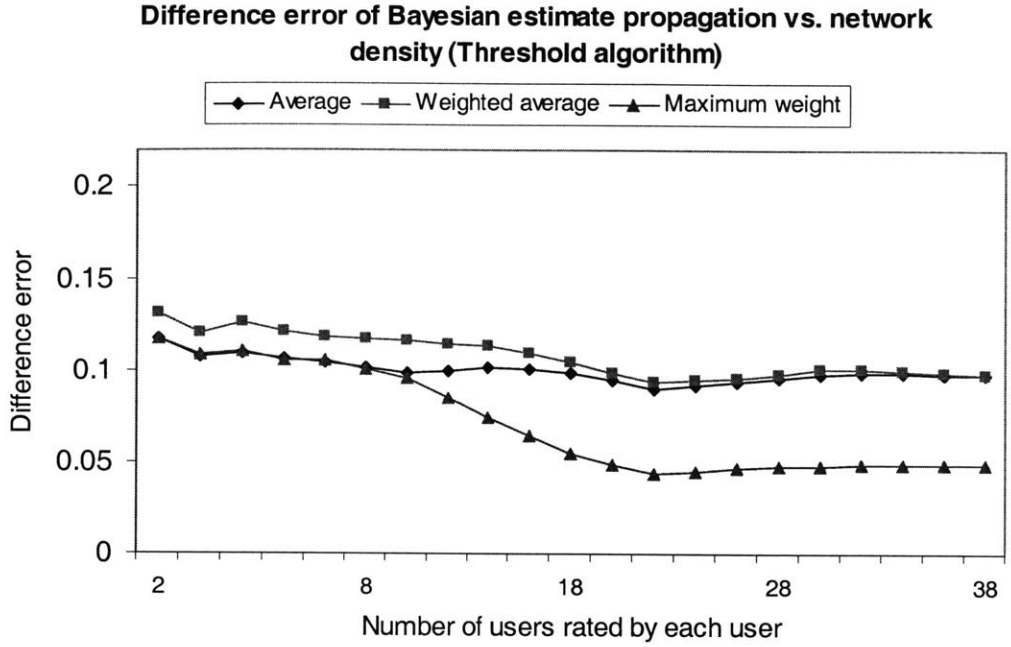


Figure 4.15. Difference error versus network density for the 3 multiple-paths inference strategies using Bayesian estimate rating propagation. The direct user-to-user rating algorithm used is the Threshold algorithm. Network density refers to number of edges over number of user nodes in a social network

4.4.4.2 Restaurant Bayesian Estimate Propagation: Multiple Paths

Shown in Figure 4.15 are the first set of simulation results for restaurant ratings using simulated users and their rating values. In this and the subsequent experiments, 40 users are simulated; the number of other users that each user rates is within the range [2, 39].

For this experiment, we created a simulation with 300 restaurants, where each user rates 100 restaurants. Figure 4.15 shows the performance of each multiple path inference strategy as the network density varies – where network density refers to the number of edges over the number of user nodes in a social network. Figure 4.15 shows that the performances of the 3 different aggregation strategies generally increase as network density increases – *i.e.*, the difference error decreases. In the case of the simulations in Figure 4.15, the best performing aggregation strategy is the maximum weight strategy (*c.f.*, Equation 4.8). This observation is especially true as the network density increases. The maximum weight strategy seems to be able to make use of the increased number of paths between pairs of users to produce the most accurate estimated ratings among the 3 strategies studied.

The same Bayesian estimate propagation scheme is applied for the 3 multiple paths aggregation strategies using the Agreement Likelihood direct user-to-user rating algorithm. Recall that for every network density value, 10 simulations runs are executed

for each strategy and an average error measure for the strategy is calculated and plotted. The results of this experiment are shown in Figure 4.16. When the network density is low, the error measure is high for all the strategies because there are very few paths linking any two pairs of users together. This limited number of paths did not allow any of the multiple-paths inference strategies to calculate accurate estimated ratings.

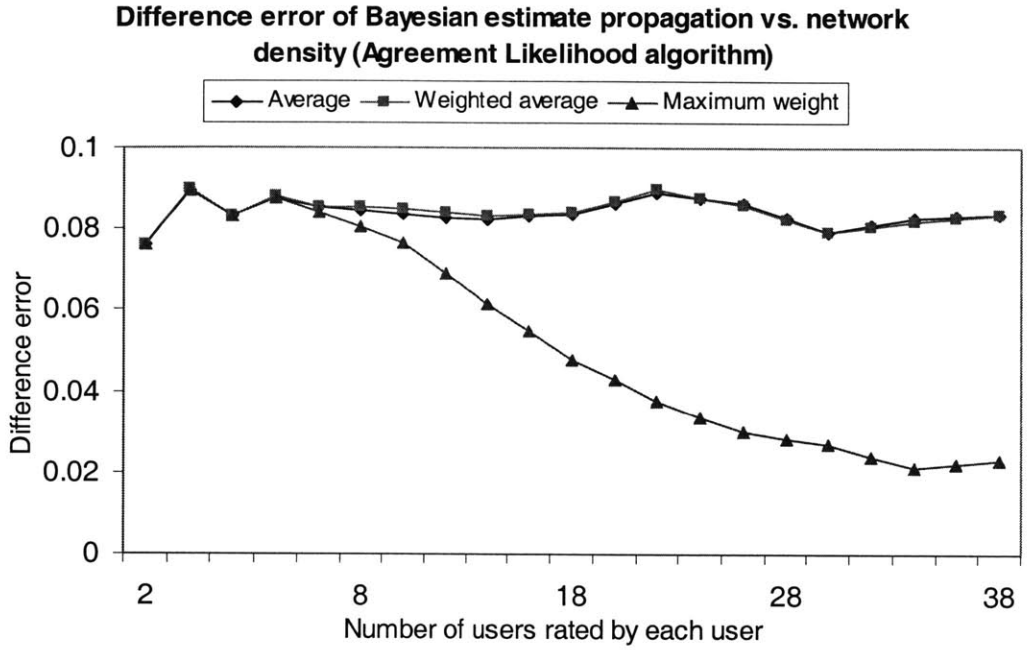


Figure 4.16. Difference error versus network density for the 3 multiple-paths inference strategies using Bayesian estimate rating propagation. The direct user-to-user rating algorithm used is the Agreement Likelihood algorithm. Network density refers to number of edges over number of user nodes in a social network

By using the Agreement Likelihood algorithm for direct user-to-user rating, Figure 4.16 further exhibits the performance of the maximum weighting multiple paths inference algorithm versus the other two being examined. When the network is sparse, the error measure used in this experiment is high; for the maximum weighting strategy, its error measures are about at the level for the other 2 strategies. However, when the network density increases, the maximum weight strategy is able to make use of the increased number of paths between pairs of users to produce the most accurate estimated ratings among the 3 strategies studied.

4.4.4.3 Restaurant Preference-based Propagation: Multiple Paths

The same experimental setup as the Bayesian estimate propagation scheme is applied with the preference-based propagation scheme. Recall that for every network density value, 10 simulation runs are executed for each strategy and an average error measure for the strategy is calculated and plotted. The results of this experiment are shown in Figure 4.17. Unlike the Bayesian estimate propagation schemes, the difference error does not diminish as much when the network density increases. Overall, the weighted average strategy for inferring rating across multiple paths is still performing not as well as the other two strategies.

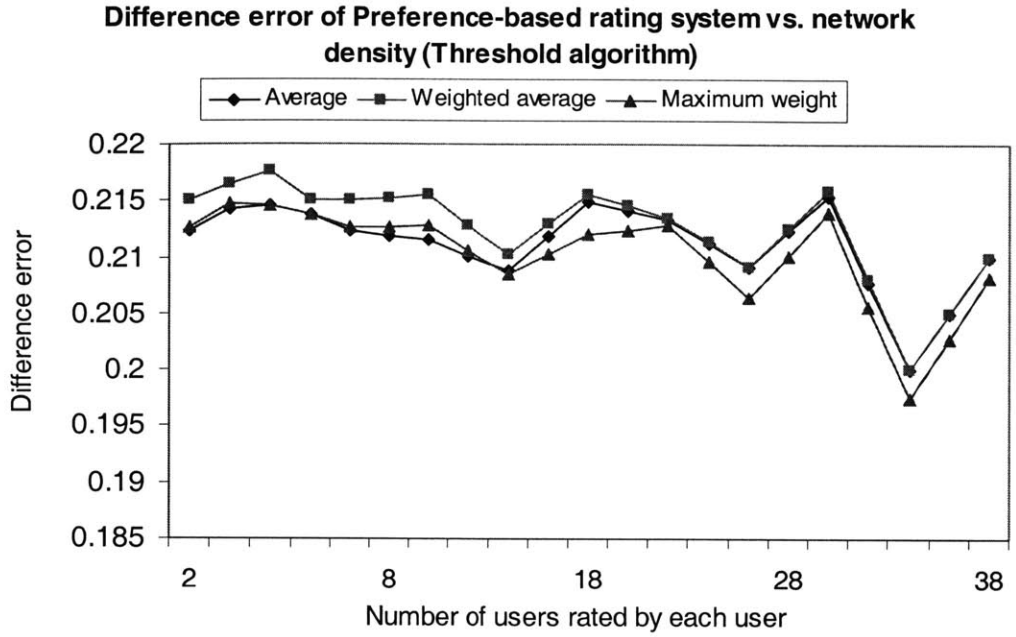


Figure 4.17. Difference error versus network density for the 3 multiple-paths inference strategies using Preference-based rating propagation. The direct user-to-user rating algorithm used is the Threshold algorithm. Network density refers to number of edges over number of user nodes in a social network

Shown in Figure 4.18 are results from a Preference-based rating propagation experiment with the simulated restaurant rating experiment. The direct user-to-user rating algorithm used is the Agreement Likelihood algorithm. Compared to the Threshold direct rating algorithm, the Agreement Likelihood algorithm generates lower difference errors for all values of direct neighborhood size in the range [2, 39]..

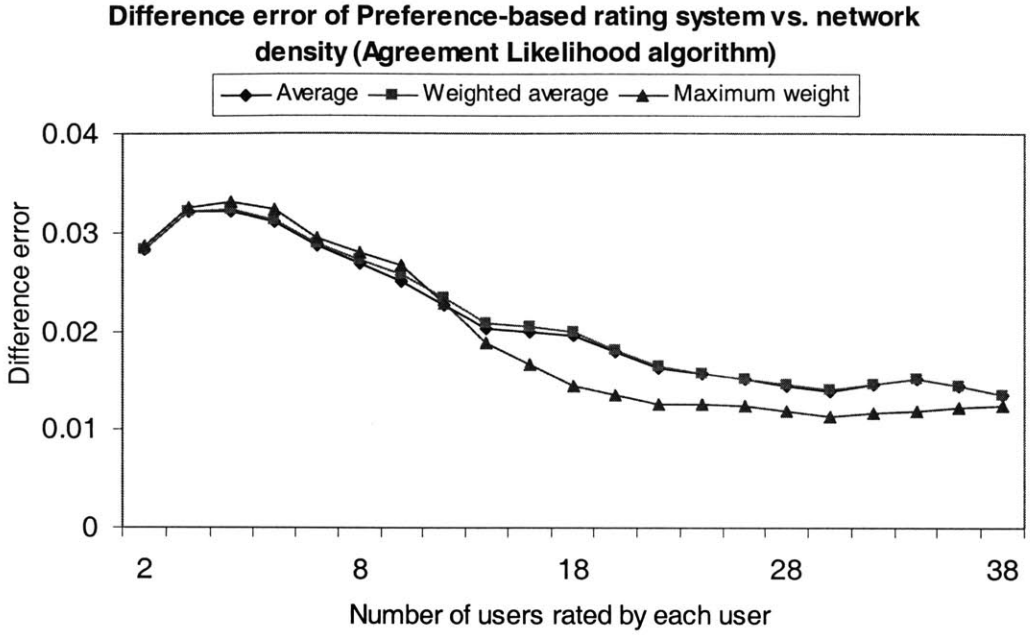


Figure 4.18. Difference error versus network density for the 3 multiple-paths inference strategies using the Preference-based rating propagation. The direct user-to-user rating algorithm used is the Agreement Likelihood algorithm. Network density refers to number of edges over number of user nodes in a social network

Overall, the performances of the 3 aggregation strategies for multiple paths are similar through this last set of multi-paths experiments. The error measure (difference error) for each one of them decreases as the network density increases until each user rates about half of the total number of users in the network. Beyond that, the error stabilizes at a steady value. The maximum weighting strategy for rating inference across multiple paths consistently performs at or above the performance of the other (averaging and weighted averaging) strategies examined.

4.5 Conclusion and Discussions

For distributed systems at large and e-commerce systems in particular, ratings play an increasingly important role. Ratings confer resource reliability or reputation of sources. The previous and this chapter have reported our formalization of the rating process and a set of experiments testing various aspects of the rating process. We have argued that since ratings and reputations are clearly subjective quantities, models about them should explicitly model their contextualized and personalized nature.

Our experimental framework makes use of a social network based on the rating patterns of a community of users. Each social network is constructed using user ratings for a given context. In our simulations, our personalized rating schemes perform about as well as that of the global reputation scheme when the network density of a social network is low. (Recall that network density refers to the ratio of number of edges over number of nodes). As the network density increases, personalized rating systems based on our

Bayesian estimate or Preference-based rating propagation schemes can significantly outperform a global rating scheme. Increase in the network density translates to increase in the amount of information about users in the social network. This observation suggests that personalized rating schemes are not always better than global rating schemes; their advantages are only realized when there are sufficient ratings in the community. In our experimental settings, the level of "sufficiency" has been tested and reported earlier in this chapter.

In comparing the Bayesian estimate and the Preference-based rating propagation schemes, our results show that in cases where the rating of resources is binary such as in the Threshold algorithm, the Bayesian estimate scheme performs better than the Preference-based scheme. This observation can be explained by noting that the Bayesian estimate scheme is modeled after binary ratings (*c.f.*, Section 3.6). In the case where ratings are values in the real numbers, as in the case of the Agreement Likelihood rating algorithm, the Preference-based rating system can outperform the Bayesian estimate scheme.

Emerging computing paradigms such as web services, peer-to-peer networking, and pervasive computing are making available a myriad of resources to users. Multiple businesses may compete in the same space by offering similar services. Descriptions of businesses and their services allow the consumer to locate a business providing an appropriate service. The description, however, does not fully indicate the quality or reliability of the services provided by the business to the consumers. Provided careful designs are made to prevent attacks, personalized rating systems as proposed here could be used in this context to provide a reputation system for businesses and their services.

There are several issues yet to be resolved. Already pointed out is the multiple paths inference problem. The rest will be discussed in the future works section in Chapter 9. This chapter has experimented with several strategies for inference in such setting. We have shown that a good strategy is to use the path that contains the most trusted intermediary. Recent work by Murphy, *et al* (1999), Yedidia, *et al.* (2001) and others have pointed to stochastic techniques for dealing with this multiple paths (or loopy networks) inference problem. Along with this issue, areas for future rating work are discussed in Chapter 9.

CHAPTER 5

A Computational Model of Trust and Reputation

Much parallel exists between the trust and reputation models in this chapter and the ratings work presented in Chapter 3 and 4. This is not surprising since ratings confer perceptions of reliability, leading to the production (or *un*-production) of trust for those viewing the ratings. One can argue that a sound rating system should be based on in-depth understanding of how trust and reputation work in a society of agents. From this view, this chapter is complementary to the ratings work in the previous two chapters. In fact, the personalized rating systems proposed in those two chapters have been enhanced as a result of the models formulated here.

Let's review the main critiques from the trust and reputation literature reviewed so far:

- Differentiation of trust and reputation is either not made or the mechanism for inference between them is not explicit.
- Trust and reputation are taken to be the same across multiple contexts or are treated as uniform across time.
- Despite the strong sociological foundation for the concepts of trust and reputation, existing computational models for them are often not grounded on understood social characteristics of these quantities.

The rest of this chapter aims to construct a computational model of trust and reputation that addresses these points. In Section 5.1, we first provide rationales behind the quantitative model proposed in this chapter for trust and reputation. We describe desirable criteria for a sociologically grounded model that addresses the inadequacies of existing models. Section 5.2 explains the notations to be used. Section 5.3 presents our computational model of reputation and related quantities. Section 5.4 provides a brief conclusion and introduction to the experiments in the next chapter evaluating our proposed model and several others from the literature.

5.1 Model Rationales

Contrary to game theorists' assumptions that individuals are rational economic agents [1] who use backward induction to maximize private utilities (Fudenberg and Tirole, 1996; Binmore, 1997), field studies show that individuals are at best boundedly rational [2] (Simon, 1996) and do not use backward induction in selecting actions [3] (Rapoport, 1997; Hardin, 1997). Social-biologists and psychologists have shown in field studies that human subjects can effectively learn and use heuristics [4] in decision making (Barkow, *et al.*, 1992; Guth and Kliemt, 1996; Trivers, 1971). One important heuristics that has been found to pervade human societies is the *reciprocity norm* for repeated interactions with the same parties (Becker, 1990; Gouldner, 1960). In fact, people use reciprocity norms even in very short time-horizon interactions (McCabe, *et al.*, 1996). *Reciprocity norms* refer to social strategies that individuals learn which prompt them to "... react to the positive actions of others with positives responses and the negative actions of others with negative responses" (Ostrom, 1998). From common experience, we know that the degree to which reciprocity is expected and used is highly variable from one individual to another. Learning the degree to which reciprocity is expected can be posed as a trust estimation problem.

Reciprocity in the context of evolutionary biology and game theory will be reviewed in detail in Chapter 7. Here, we highlight a few important points to further the rationales for our trust and reputation models being proposed. There are many reciprocity strategies proposed by game-theoreticians; the most well-known of which is the tit-for-tat strategy, which has been extensively studied in the context of the Prisoners's Dilemma game (Axelrod, 1984; Pollock and Dugatkin, 1992; Nowak and Sigmund, 2000). Not everyone in a society learns the same norms in all situations. Structural variables affect individuals' level of confidence and willingness to reciprocate. In the case of cooperation, some cooperate only in contexts where they expect reciprocation from their interacting parties. Others will only do so when they are publicly committed to an agreement.

When facing social dilemmas [5], trustworthy individuals tend to **trust** others with a reputation for being trustworthy and shun those deemed less so (Cosmides and Tooby, 1992). In an environment where individuals "regularly" perform **reciprocity** norms, there is an incentive to acquire a **reputation** for reciprocative actions (Kreps, 1990; Milgrom, *et al.*, 1990; Ostrom, 1998). "Regularly" refers to a *caveat* observed by sociologists that reputation only serves a normative function in improving the fitness of those who cooperate while disciplining those who defect if the environment encourages

---

[1] Rational agents refer to those able to deliberate, ad infinitum, the best choice (for maximizing their private utility functions) without regard to computational limitations (*c.f.*, Fudenberg and Tirole, 1991).

[2] Bounded rationality refers to rationality up to limited computational capabilities (*c.f.*, Simon, 1981)

[3] Backward induction here refers to a style of inference based on inducting from the last game of a sequence of games by maximizing a given utility at each step (this style can also be characterized as dynamic programming) (*c.f.*, Axelrod, 1984; Fudenberg and Tirole, 1996).

[4] A heuristic refers to "rules of thumb — that [individuals] have learned over time regarding responses that tend to give them good outcomes in particular kinds of situations." (Ostrom, 1998).

[5] Social dilemma refers to a class of sociological situations where maximization of personal utilities do not necessarily lead to the most desirable outcome. Tragedy of the commons (Hardin, 1968) or Prisoner's dilemma (Axelrod, 1984) is the most studied social dilemma.

the spreading of reputation information (Castelfranchi, *et al.*, 1998). In the words of evolutionary biologists, having a good reputation increases an agent's *fitness* in an environment where reciprocity norms are expected (Nowak and Sigmund, 1998). Therefore, developing the quality for being trustworthy is an asset since trust affects how willing other individuals are to participate in reciprocative interactions (Dasgupta, 2000; Tadelis, 1999).

The following section will transform these statements into mathematical expressions. The formulation below is very similar to the Bayesian estimate and Preference-based rating propagation schemes proposed in Chapter 3.

The intuition behind the model given here is inspired by Ostrom's 1998 Presidential Speech to the American Political Society, which proposed a qualitative behavioral model for collective action. Section 2.2 briefly discussed Ostrom's work in the context of research on reputation across multiple fields.

To facilitate the model description, agents and their environment are to be defined. Consider the scenario that agent a_j is evaluating a_i's reputation for being cooperative. The set of all agents that a_j asks for this evaluation can be considered to be a unique society of N agents **A** (where both the elements in **A** and its size depend on different a_j's). **A** is called an "embedded social network" with respect to a_j (Granovetter, 1985):

- **Agents**: $\mathbf{A} = \{a_1, a_2, \ldots a_N\}$

Clearly, a_i must be part of **A** in order for a_j to evaluate a_i in a non-trivial manner. The reputation of an agent a_i is *relative* to the particular embedded social network in which a_i is being evaluated.

It should be clear from the argument thus far that reciprocity, trust and reputation are highly related concepts. The following relationships are expected:

- Increase in agent a_i's reputation in a_j's embedded social network **A** should also increase the trust from a_j for a_i (where a_j can be any agent that "knows" a_i. *i.e.*, a_j is in the embedded social network of a_i)
- Increase in an agent a_j's trust of a_i should also increase the likelihood that a_j will reciprocate positively to a_i's action.
- Increase in a_i's reciprocating actions to other agents should also increase a_i's reputation in the embedded social network of a_j.
- Decrease in any of the three variables should lead to the reverse effects.

Graphically, these intuitive statements create the relationships among the three variables of interest as shown in Figure 5.1.

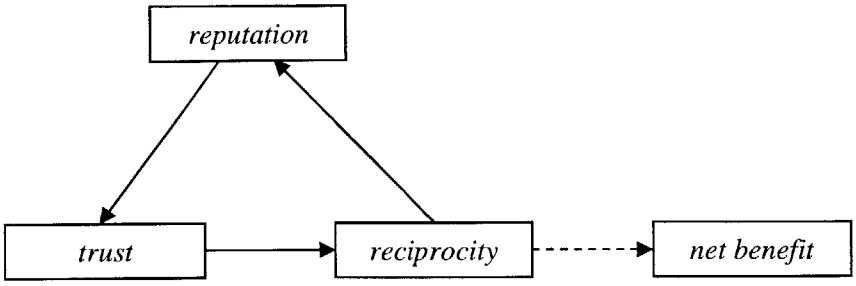


Figure 5.1. This simple model shows the reinforcing relationships among trust, reputation and reciprocity. The direction of the arrow indicates the direction of influence among the variables. The dashed line indicates a mechanism not discussed.

5.1.1 Reciprocity

This chapter uses the following definition for reciprocity:

- **Reciprocity**: mutual exchange of deeds (such as favor or revenge) [6]

This definition is largely motivated by the many studies of reciprocity in which repeated games are played between two or more individuals (Raub and Weesie, 1990; Boyd and Richersen 1989; Nowak and Sigmund, 1998; *c.f.*, Chapter 7). Two types of reciprocity are considered: direct reciprocity refers to interchange between two concerned agents. Indirect reciprocity refers to actions where the initiator of an action does not obtain a returned action by a recipient; rather, the reaction from the recipient is fed back to the initiator through a third party. For example, in an analogy for donor of good deeds, the donor might not necessarily be rewarded by the recipient directly but by other individuals who might be recipients of other good deeds by others.

Reciprocity can be measured in two ways. Firstly, reciprocity can be viewed as a social norm shared by agents in a society. The higher this "societal reciprocity," the more likely one expects a randomly selected agent from that society to engage in reciprocating actions. Secondly, reciprocity can be viewed as a dyadic variable between two agents (say a_i and a_j). The higher this "dyadic reciprocity," the more one expects a_i and a_j to reciprocate each other's actions. In this latter case, no expectation about other agents should be conveyed. For any single agent a_i, the cumulative dyadic reciprocity that a_i engages in with other agents in a society should have an influence on a_i's *reputation* as a reciprocating agent in that society.

5.1.2 Reputation

Much of the literature on reputation has been reviewed in Chapter 2. For our work, we use the following definition for reputation:

- **Reputation**: perception that an agent has of another's intentions and norms [7]

Reputation is a social quantity calculated based on actions by a given agent a_i and observations made by others in an "embedded social network" in which a_i resides

[6] Ostrom (1998) further discusses how reciprocity affects the level of cooperation which affects the overall net benefits in a society.

[7] Ostrom (1998) defines norm as "... heuristics that individuals adopt from a moral perspective, in that these are the kinds of actions they wish to follow in living their life."

(Granovetter, 1985). a_i's reputation clearly affects the amount of trust that others have toward it. Now, how is trust defined?

5.1.3 Trust

The definition for trust by Gambetta (1988) is often quoted in the literature: "...trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both *before* he can monitor such action (or independently of his capacity ever to be able to monitor it) *and* in a context in which it affects *his own* action" (*ibid.*). We elect the term "subjective expectation" rather than "subjective probability" to emphasize the point that trust is a summary quantity that an agent has toward another based on *a number of former encounters* between them:

- **Trust**: a subjective expectation an agent has about another's future behavior.

Trust is a subjective quantity calculated based on the two agents concerned in a present or future dyadic encounter. Dasgupta (2000) gave a similar definition for trust: the expectation of one person about the actions of others that affects the first person's choice, when an action must be taken before the actions of others are known.

Given the simple model of interaction in Figure 5.1, the rest of this chapter operationalizes this model into mathematical statements that can be implemented in a real world system.

5.2 Notations

To simplify the reasoning about the main quantities of interest (reciprocity, trust, and reputation), two simplifications are made here. First, the embedded social networks in which agents are embedded are taken to be static. *i.e.*, no new agents are expected to join or leave. Secondly, the action space is restricted to be:

- **Action**: $\alpha \in$ { cooperate, defect }

In other words, only binary actions are considered. Let $0 < \gamma < 1$ represents the amount of reciprocity in the embedded social network where low γ represents low level of reciprocity and *vice versa*:

- **Reciprocity**: $\gamma \in [0, 1]$

γ measures the fraction of reciprocative actions that occur in a society. In other words, "*cooperate*" actions are met with "*cooperate*" response; "*defect*" actions are met with "*defect*" responses. How γ is derived in our model will be discussed shortly.

Let **C** be the set of all contexts of interest. The reputation of an agent is a social quantity that varies with time. Let $\theta_{ji}(c)$ represent a_i's reputation in an embedded social network of concern to a_j for the context $c \in \mathbf{C}$. In this sense, reputation for a_i is subjective to every other agent since the embedded social network that connects a_i and a_j is different for every different a_j. Reputation is the perception that suggests an agent's intentions and norms in the embedded social network that connects a_i and a_j. $\theta_{ji}(c)$ measures the likelihood that a_i reciprocates a_j's actions, and can be reasonably represented by a probability measure:

- **Reputation**: $\theta_{ji}(c) \in [0, 1]$

Low $\theta_{ji}(c)$ values confer low intention to reciprocate and high values indicate otherwise. As agent a_i interacts with a_j, the quantity $\theta_{ji}(c)$ as estimated by a_j is updated with time as a_j's perception about a_i changes.

To model interactions among agents, the concept of an encounter between two agents is necessary. An encounter is an event between two agents (a_i, a_j) within a specific context such that a_i performs action α_i and a_j performs action α_j. Let **E** represent the set of encounters. This set is characterized by:

- **Encounter**: $e \in \mathbf{E} = \alpha^2 \times \mathbf{C} \cup \{ \perp \}$

where $\{\perp\}$ represents the set of no encounter ("bottom"). While evaluating the trustworthiness of a_i, any evaluating agent a_j relies on its knowledge about a_i garnered from former encounters or hearsay about a_i. Let $D_{ji}(c)$ represents a history of encounters that a_j has with a_i within the context c:

- **History**: $D_{ji}(c) = \{\mathbf{E}^*\}$

where * represents the Kleene closure, and D_{ji} might include observed encounters involving other agents' encounters with a_i. Based on $D_{ji}(c)$, a_j can calculate its trust toward a_i, which expresses a_j's expectation of a_i's intention for reciprocation. The above statement can be translated to a pseudo-mathematical expression (which is explained latter in the chapter):

- **Trust**: $\tau(c) = \mathrm{E}[\, \theta(c) \mid D(c) \,]$

The higher the trust level for agent a_i, the higher the expectation that a_i will reciprocate agent a_j's actions.

5.3 Computational Models

Consider two agents a and b, who care about each others' actions with respect to a specific context c. For clarity, a single context 'c' is used for all variables. We are interested to have an estimate for b's reputation in the eyes of a: θ_{ab}. Here we assume that a always perform "cooperate" actions and that a is assessing b's tendency to reciprocate cooperative actions. Let a binary random variable $x_{ab}(i)$ represent the ith encounter between a and b. $x_{ab}(i)$ takes on the value '1' if b's action is 'cooperate' (with a) and '0' otherwise. Let the set of n previous encounters between a and b be represented by: [8]

- **History**: $D_{ab} = \{ x_{ab}(1), x_{ab}(2), \ldots, x_{ab}(n) \}$

Let p be the number of cooperative events by agent b toward a in the n previous encounters. b's reputation θ_{ab} for agent a should be a function of both p and n. A simple function can be the proportion of cooperative action over all n encounters. From statistics, a *proportion* random variable can be modeled as a Beta distribution (Dudewicz and Mishra, 1988): $\Pr(\hat{\theta}) = Beta(c_1, c_2)$ where $\hat{\theta}$ represents an estimator for θ, and c_1 and

[8] For clarify, the discussion takes the viewpoint of "direct" encounters between a and b. It is equally sensible to include observed encounters about a's actions toward others.

c_2 are parameters determined by prior assumptions — as discussed later in this section. This proportion of cooperation in n finite encounters becomes a simple estimator for θ_{ab}:

$$\hat{\theta}_{ab} = \frac{p}{n}$$

Assuming that each encounter's cooperation probability is independent of other encounters between a and b, the likelihood of p cooperations and $(n - p)$ defections can be modeled as: $L(D_{ab} \mid \hat{\theta}) = \hat{\theta}^p (1-\hat{\theta})^{n-p}$. The Beta distribution turns out to be the conjugate prior for this likelihood (Heckerman, 1996). Combining the prior and the likelihood, the posterior estimate for $\hat{\theta}$ becomes (the subscripts are omitted):

$$\Pr(\hat{\theta} \mid D) = Beta(c_1 + p, c_2 + n - p)$$

The steps of derivation for this formula are given in (*Mui, et al. 2001*). First order statistical properties of the posterior are summarized below for the posterior estimate of $\hat{\theta}$:

$$\mathrm{E}\left[\hat{\theta} \mid \mathrm{D}\right] = \frac{c_1 + p}{c_1 + c_2 + n} \qquad \sigma_{\hat{\theta}|D}^{\ 2} = \frac{(c_1 + p)(c_2 + n - p)}{(c_1 + c_2 + n - 1)(c_1 + c_2 + n)^2}.$$

In their next encounter, a's estimate of the probability that b will cooperate can be shown to be (*ibid.*):

$$\tau_{ab} = Pr(\, \mathrm{x_{ab}}(n+1) = 1 \mid \mathrm{D}\,) = \mathrm{E}\,[\,\hat{\theta} \mid \mathrm{D}\,]$$

Based on our model shown in Figure 5.1, **trust** toward b from a is this conditional expectation of $\hat{\theta}$ given D. The following theorem provides a bound on the parameter estimate $\hat{\theta}$.

Theorem *(Chernoff Bound).* Let $x_{ab}(1)$, $x_{ab}(2)$, ... $x_{ab}(m)$ be a sequence of m independent Bernoulli trials [9], each with probability of success $\mathrm{E}(x_{ab}) = \theta$. Define the following estimator:

$$\hat{\theta} = \left(x_{ab}(1) + x_{ab}(2) + \cdots + x_{ab}(m)\right)/m$$

$\hat{\theta}$ is a random variable representing the portion of success, so $\mathrm{E}[\hat{\theta}] = \theta$. Then for $0 \le \varepsilon \le 1$ and $0 \le \delta \le 1$, the following bound hold:

$$\Pr\left[\left|\theta - \hat{\theta}\right| \ge \varepsilon\right] \le 2\mathrm{e}^{-2\mathrm{m}\varepsilon^2} \le \delta \qquad \square$$

The proof is a straightforward application of the additive form of the Chernoff (Hoeffding) Bound for Bernoulli trials (Ross, 1995). Note that "success" in the theorem refers to cooperation in our example, but to reciprocation in general. Also note that ε refers to the deviation of the estimator from the actual parameter. In this sense, ε can be considered as a fixed *error* parameter (*e.g.*, 0.05).

From the theorem, m represents the minimum number of encounters necessary to achieve the desired level of confidence and error. This minimum bound can be calculated as follows:

---

[9] The independent Bernoulli assumption made here for the sequence of encounters is unrealistic for repeated interactions between two agents. Refinements based on removing this assumption are work in progress.

$$m \geq -\frac{1}{2\varepsilon^2}\ln\left(\delta/2\right)$$

Let $\gamma_c = 1-\delta$. γ_c is a confidence measure on the estimate $\hat{\theta}$. A γ_c approaches 1, a larger m is required to achieve a given level of error bound ε. γ_c can be chosen exogenously to indicate an agent's level of confidence for the estimated parameters.

In our model, **reciprocity** represents a measure of reciprocative actions among agents. A sensible measure for "dyadic reciprocity" is the proportion of the total number of cooperation/cooperation and defection/defection actions over all encounters between two agents. Similarly, "societal reciprocity" can be expressed as the proportion of the total number of cooperation/cooperation and defection/defection actions over all encounters in a social network. All encounters are assumed to be dyadic; encounters involving more than two agents are not modeled.

Let γ_{ab} represent the measured dyadic reciprocity between agent a and b. If $\gamma_{ab} < \gamma_c$, calculated reputation and trust estimates fall below the exogenously determined critical value γ_c and are not reliable.

5.3.1 Complete Stranger Prior Assumption

If agents a and b are complete strangers — with no previous encounters and no mutually known friends, an ignorance assumption is made. When these two strangers first meet, their estimate for each other's reputation is assumed to be uniformly distributed across the reputation's domain:

$$\Pr(\hat{\theta}) = \begin{cases} 1 & 0 < \hat{\theta} < 1 \\ 0 & \text{otherwise} \end{cases}$$

For the Beta prior, values of c_1=1 and c_2=1 yields such a uniform distribution.

5.3.2 Mechanisms for Inferring Reputation

The last section has considered how reputation can be determined when two agents are concerned. This section extends the analysis to arbitrary number of agents.

5.3.2.1 Parallel Network of Acquaintances

A schematic diagram of an embedded parallel social network for agents a and b is shown in Figure 5.2.[10]

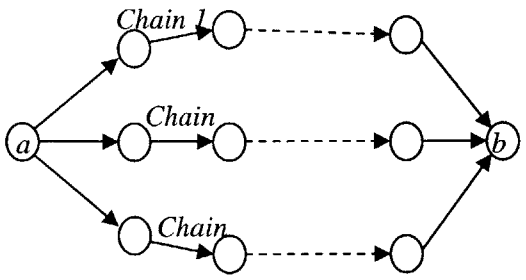


Figure 5.2. Illustration of a parallel network between two agents a and b.

[10] "Embedded social network" refers to the earlier discussion in Section 3.

Figure 5.2 shows a parallel network of k chains between two agents of interest, where each chain consists of at least one link. Agent a would like to estimate agent b's reputation as defined by the embedded network between them.[11] Clearly, to combine the parallel evidence about b, measures of "reliability" are required to weight all the evidence.

From the last section, a threshold (m) can be set on the number of encounters between agents such that a reliability measure can be established as follows:

$$w_{ij} = \begin{cases} \frac{m_{ij}}{m} & \text{if } m_{ij} < m \\ 1 & \text{otherwise} \end{cases}$$

where m_{ij} is the number of encounters between agents i and j. The intuition behind this formula is as follows: arguments by Chernoff bound in the last section have established a formula to calculate the minimum sample size of encounters to reach a confidence (and error) level about the estimators. Above a given level of sample size, the estimator is guaranteed to yield the specified level of confidence. Therefore, such an estimate can be considered as "reliable" with respect to the confidence specification. Any sample size less than the threshold m is bound to yield less reliable estimates. As a first order approximation, a linear drop-off in reliability is assumed here.

For each chain in the parallel network, how should the total weight be tallied? Two possible methods are plausible: additive and multiplicative. The problem with additive weight is that if the chain is "broken" by a highly unreliable link, the effect of that unreliability is local to the immediate agents around it. In a long social chain however, an unreliability chain is certain to cast serious doubt on the reliability of any estimate taken from the chain as a whole. On the other hand, a multiplicative weighting has "long-distance" effects in that an unreliable link affects any estimate based on a path crossing that link. The form of a multiplicative estimate for chain i's weight (w_i) can be:

$$w_i = \prod_{j=1}^{l_i} w_{ij} \quad \text{where } 0 \le i \le k$$

where l_i refers to the total number of edges in chain i and w_{ij} refers to the j^{th} segment of the i^{th} chain.

Once the weights of all chains of the parallel network between the two end nodes are calculated, the estimate across the whole parallel network can be sensibly expressed as a weighted sum across all the chains:

$$r_{ab} = \sum_{i=1}^{k} r_{ab}(i)\overline{w_i}$$

where $r_{ab}(i)$ is a's estimate of b'reputation using path i and $\overline{w_i}$ is the normalized weight of path i ($\overline{w_i}$ sum over all i yields 1). r_{ab} can be interpreted as the overall perception that a garnered about b using all paths connecting the two.

[11] In general, embedded social networks do not form non-overlapping parallel chains and they are rather arbitrary (see section 5.2).

5.3.2.2 Generalized Network of Acquaintances

A schematic diagram of an arbitrary social network between agents *a* and *b* is shown in Figure 5.3.

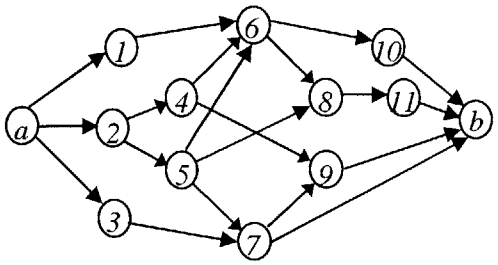


Figure 5.3. Illustration of a generalized network between two agents *a* and *b*.

If the social network in Figure 5.3 is treated as a Bayesian network, each nodes has the task of combining incoming evidence and outputting some aggregation of the inputs. In general, this pure Bayesian approach creates a parameter space that is exponential in the number of nodes (Castillo, *et al.*, 1997). This is not computationally desirable for real world systems. The most common work-around is to assume some type of causal independence and assume an aggregation technique known as "noisy-OR" (Pearl, 1988). The variables (indicating reputation for the agent being pointed at) are not independent. Work by Diez (1993; 1996) to generalize the noisy-OR to include multiply connected networks relies on assumptions about the statistics of the variables, and "strong assumptions of independence" (Diez, 1993).

Due to the difficulties raised above on the use of the noisy-OR and variant techniques, we resort to a statistical significance approach to the graph in Figure 5.3.

Given ε, δ, and consequently a minimum measure of reliability m, a graph transformation algorithm can be applied to a generalized network to reduce it to a parallel network (*c.f.*, Mui and Mohtashemi, 2002):

Algorithm (Graph Parallelization):

- For every node i in the network, define $I(i)$ as the indegree of i and $O(i)$ as the outdegree of i.
- If for all nodes other than the source and the sink $I(i) = O(i) = 1$, then the graph must be a parallel network. Proceed as in the previous section.
- Otherwise for each node i, with $I(i), O(i) > 1$, look up the number of encounters for each one of its $I(i) + O(i)$ direct links.
- For every node i, with $I(i), O(i) > 1$, remove those links with reliability below a threshold $t < m$. t is application dependent and is a function of both the size of the network and the amount of error the investigator is willing to tolerate.
- For every node i that after step 4 still has $I(i), O(i) > 1$, form as many as $I(i) \times O(i)$ parallel paths each through a copy of node i. The new graph must be a parallel network.

Let's apply the above algorithm to the example in Figure 5.3: $O(2)=O(4)=O(6)=O(7)=I(7)=I(8)=I(9)=2$, and $O(5)=I(6)=3$. Suppose further that the reliability of the links 6→8 and 5→7 is below the threshold t. Then the network depicted in Figure 3 can be transformed into the parallel network show in Figure 5.4.

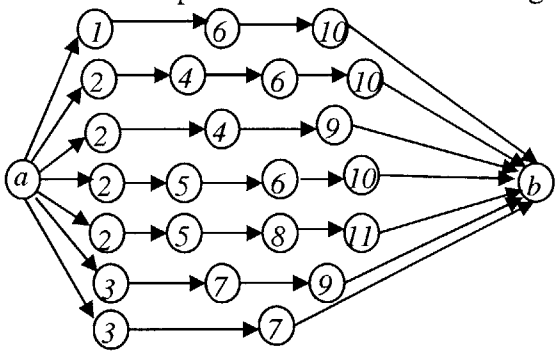


Figure 5.4. Illustrated is a parallel network resulting from the application of the graph parallelization algorithm to the example of Figure 5.3.

Once a generalized network is reduced into a parallel network, the steps in section 5.1 can be followed for calculating reputation related quantities discussed in this chapter.

5.4 Discussions

No model is perfect for social phenomena. Our proposed models for trust and reputation have integrated much of what has been learned in diverse fields on these social quantities. In this section, we outline a few questions that might arise as a result of the models presented here. We attempt to resolve as many of these questions as we are able to within the realms justified by our models.

5.4.1 The Ghandi or Christ Question

Based on our models in this chapter, one can raise the following question:

"Would Mahatma Ghandi (or Jesus Christ) get a lower reputation because he tends to err on the side of cooperation even when they 'should' defect?"

The underlying concern for this question is about the mechanism for reciprocity. The questioner has in mind reciprocity in the form of a globally defined tit-for-tat strategy based on an action space with two actions.

In the context of our models, this question is flawed.

It relies on the assumption of a universally acknowledged action space and a universal understanding of the notion of reciprocity. Our trust and reputation models are based on the notion of "embedded social network" (*c.f.* Section 5.1) where each social network is defined with respect to an agent evaluating the network. To recapitulate, the reputation (or levels of trust or reciprocity) of an agent a_i is *relative* to the particular embedded social network in which a_i is being evaluated. When being betrayed, Mahatma Ghandi (or Jesus Christ) still elects to cooperate. This action is in the spirit of: "If someone strikes you on one cheek, turn to him the other also."[12] *Turn the other cheek* is the proper *reciprocity* to those who subscribe to Ghandi or Christ's teachings.

In addition to how actions can be perceived differently, the actions themselves can also be viewed as being relative with respect to the agent exercising them. A similar argument follows. If cooperation among mankind is what they are striving after, Ghandi or Christ might very well consider "turn the other cheek" type of actions to be promoting the *right kind* of reciprocity, as opposed to the *earthly* kind.

5.4.2 The Einstein Problem

From the above discussion, one might infer that our trust and reputation models are based on the approval of one toward another about a specific context. One can raise another objection: most people would consider Einstein as a very reputed physicist but most would not be in the position to "approve" Einstein on his General Theory of Relativity.

The concept of delegation helps in alleviating this conceptual difficulty. Most people know someone who knows someone who knows others who are physicists. These physicists seem all to approve Einstein on his works. Those who know these physicists agree that these physicists are physicists and they therefore delegate the reputation evaluation for Einstein as a physicist to them. By delegation, the non-physicists approve of Einstein.

[12] From the Gospel of Luke 6:29.

Consider the following scenario: *a* is not a physicist but knows a friend *b* who is a physicist. *a* does not know any other individual personally and does not trust any other people nor media. *Figure 2* captures this situation:

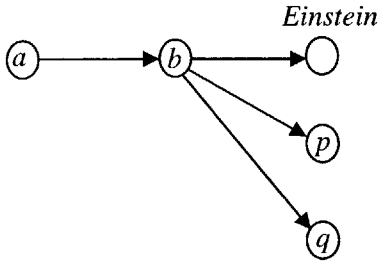


Figure 5.5. This graph illustrates approval delegation. *b* (a physicist) approves of Einstein as a physicist. This social network shows that *b* knows of another 2 physicists *p* and *q*. (Arrows indicates whether one knows about another's existence: through previous encounters or ratings.) If *a* approves of *b* and *b* approves of *Einstein*, then *a* should approves of *Einstein* even though *a* does not have any direct knowledge of Einstein.

Referring to *Figure 2*, for every encounter between *a* and *b* about physics, since *a* does not know anyone else who is a physicist, *a* always defer to *b*'s physics knowledge — *b* has a high reputation in physics in *a*'s mind. *b* approves of Einstein as a physicist. Therefore, *a* agrees with *b* (whom *a* approves of about physics) that Einstein is a reputed physicist.

Approval delegation can be propagated for arbitrary levels, as is the case in our proposed rating propagation scheme in Chapter 3 and 4.

5.5 Conclusion

This chapter has listed desiderata and presented a formalism that satisfies them. We have attempted to integrate our understanding across the surveyed literatures to construct a computational model of rational decision making use of social information by quantifying the notions of trust and reputation. Our model has the following characteristics:

- makes explicit the difference between trust and reputation;
- defines social information as the union of the amount of information embedded in the social structure as dictated by trust and reputation
- defines reputation as a quantity relative to the particular embedded social network of the evaluating agent and the history of encounters;
- defines trust as a dyadic quantity between the trustor and the trustee which can be inferred from social information from those in the embedded network;
- by considering all possible paths to the source of information (k-degree-acquaintance) in the reputation framework, the model effectively increases the underlying sample size for estimating trust and reputation;

- proposes a probabilistic mechanism for inference by quantifying concepts such as trust, reputation, and level of reciprocity.

The explicit formulation of trust, reputation, and related quantities suggests a straightforward implementation of the model in a multi-agent environment (such as an electronic market). Comparing one such implementation of this model against existing models is the task of the next chapter on reputation experiments. We now investigate whether our proposed formulation of trust and reputation can be compared quantitatively with existing schemes. This is the task of the experiments in the next chapter.

CHAPTER 6

Reputation Experiments

This chapter describes a set of experiments for an evolutionary game known as the iterated Prisoner's Dilemma (Maynard-Smith, 1982; Axelrod, 1984). An extended version of this simulation is discussed in Chapter 8 in the context of evolution of cooperation. Our aim is to establish a quantitative framework in which to compare the various notions of reputation that have been proposed by us and those existing in the literature.

6.1 Simulation Framework

If reputation has a utility value for the survival of an agent, we would like to design a set of experiments to test which notion of reputation provides the highest utility. We use an evolutionary version of the incomplete information game similar to that used in Kreps and Wilson (1982) and Milgrom and Roberts (1982).

Evolutionary games are made popular by Maynard-Smith (1982)'s work. Whereas iterated games are played between the same players over time, evolutionary games are groups of iterated games played across multiple "generations" of related players. Consider an evolutionary game of 1 generation, this game is exactly the iterated game scenario. An evolutionary game of 2 generations starts off with 1 generation of agents each playing iterated games with other agents. At the end of a preset number of rounds that mark the end of a generation, each agent procreates certain number of children which is a function of the accumulated fitness of the agent. After procreation, the 1^{st} generation agents are removed from the game. The set of children for all 1^{st} generation agents then form the agents for the 2^{nd} generation of games.

6.1.1 Indirect Reciprocity

Indirect reciprocity will be discussed in detail in Chapter 7. We briefly review the highlight of this concept below for our simulation.

In the field of evolutionary game theory, "evolution of cooperation" is an important research problem (Axelrod, 1984). Trivers (1971) has suggested the idea of *reciprocal altruism* as an explanation for the evolution of cooperation. Altruists indirectly contribute to their fitness (for reproduction) through others who reciprocate back.

Reputation can potentially help to distinguish altruists from those disguised as such, thereby preventing those in disguise from exploiting the altruists. Alexander (1987) greatly extended this idea to the notion of *indirect reciprocity*. In situations involving cooperators and defectors, *indirect reciprocity* refers to reciprocating toward cooperators indirectly through a third party. One important heuristic that has been found to pervade human societies is the *reciprocity norm* for repeated interactions with the same parties (Becker, 1990; Gouldner, 1960).[1] Therefore, a reasonable model for a human is an agent engaged in reciprocal interactions.

In the following sub-section, groups of reciprocating agents are simulated against all-defecting agents. By using various notions of reputation, the reciprocating strategy can be shown to be superior from the standpoint of survivability.

6.1.2 Simulation Framework

For the Prisoner's Dilemma (PD) game, the action space for each agent is:

Action = { *cooperate, defect* }

The payoff matrix for the Prisoner's Dilemma game is (where T > R > P > S and 2R > T + S. *c.f.*, Fudenberg and Tirole, 1991):

		agent 2: C	agent 2: D
agent 1	C	R, R	S, T
	D	T, S	P, P

Figure 6.1. *Payoff matrix for a one-shot Prisoner's dilemma game, where C = cooperate, D = defect. In the 4x4 cell, the letter on the left refers to agent 1's payoff; the letter on the right refers to agent 2's payoff.*

In the game-theoretic literature for describing payoffs for the Prisoner's dilemma (PD) game, four descriptions are used for the outcomes of the game: temptation, reward, punishment, and sucker:

- *Temptation (T): when an agent defects (D) while its opponent cooperates (C)*
- *Reward (R): when both an agent and its opponent cooperate (C) with each other*
- *Punishment (P): when both an agent and its opponent defect (D) against each other*
- *Sucker (S): when an agent cooperates (C) while its opponent defects (D)*

Participants in an encounter are chosen randomly from the population. After the first participant is selected, a second participant is randomly selected. At the end of a generation (where a certain number of dyadic encounters between agents have occurred), an agent begets progeny in the next generation proportional to that agent's total fitness. The total population size is fixed, so any increase in the number of one type of agent is balanced by a decrease in the numbers of other types of agents.

[1] *Reciprocity norms* refer to social strategies that individuals learn which prompt them to "... react to the positive actions of others with positives responses and the negative actions of others with negative responses (Ostrom, 1998).

6.1.3 Simulation Parameters and Agent Strategies

For each of the simulation experiments, 50 agents with strategy tit-for-tat (TFT) and 50 agents with strategy all defecting (AllD) are mixed into a shared environment. These strategies are defined below. A total of 30 generations are simulated per experimental run (during which no new agents are introduced into the system). The payoff values (*c.f.*, Figure 6.1) are: $T = 5$, $R = 3$, $P = 1$, $S = 0$.

We studied agent strategies in which the decision for an encounter with an agent is based on the last interaction with that agent. Each strategy is characterized by five probabilities for cooperation: an initial probability and four probabilities for each of the possible outcomes of the last encounter. We extended these strategies by adding a reputation threshold that determines how an agent will act. Example agent strategies for this game are:

- *Cooperate (C): always cooperates.*
- *Defect (D): always defects.*
- *Tit-for-tat (TFT): initially cooperates, and then does what the other agent did in the last round.*
- *Reputation tit-for-tat (RTFT): initially cooperates depending on the reputation of the other agent, and then does whatever the other agent did in the last round*

The reputation referred to for RTFT agents is determined using one of several reputation notions as described below. If the reputation of the target agent is less than a minimum reputation threshold, then the RTFT agent defects, otherwise it cooperates.

Strategies	**I**	**T**	**R**	**P**	**S**
Cooperate (C)	1	1	1	1	1
Defect (D)	0	0	0	0	0
Tit-for-tat (TFT)	1	1	1	0	0
Reputation Tit-for-tat (RTFT)	*	1	1	0	0

Figure 6.2 Probabilities of cooperation for different strategies. The column labeled I gives the initial probability for cooperation, while those labeled T, R, P, and S give the probabilities for cooperation given that the outcome (payoff) of the previous encounter was temptation, reward, punishment, or sucker. The initial probability for RTFT (*) depends on opponent's reputation and the reputation threshold used.

6.1.4 Goal of Simulation

In studies of evolutionary games, one phenomenon that is of interest is whether certain strategies survive after many generations of evolution. Those strategies that survive (as manifested by a viable set of progeny agents with that strategy) are called *evolutionarily stable.*

In our simulations, we studied the conditions under which TFT agents are evolutionarily stable when they use different notions of reputation to judge agents with

whom they interact. Specifically, we examined the "number of encounters per generation" (EPG) threshold for reputation-enhanced TFT (RTFT) to become the evolutionarily stable strategy (ESS, *c.f.* Maynard Smith, 1982). Reputation should aid agents more when more information about other agents' behavior is available. When no agents have met each other before, there is no information to calculate any reputation. The more encounters per generation occur, the more chances each RTFT agent has to learn the real reputation of the opponent agents. Note that each agent does not know the strategy of the other agents. Agents can only observe the behavior of other agents. Therefore, it is not true that once an agent is observed acting *defect*, it is therefore an AllD agent.

6.1.5 Notions of Reputation Simulated

Encounter-derived individual reputation r_e is simulated by having each RTFT agent remember encounters it has with every agent it has met before. Encounter-derived individual reputation is then the ratio of number of cooperations directly encountered over total number of encounters with a specific opponent. Such an RTFT agent defects if $r_e < r_c$ where r_c represents a critical threshold point of defection, which can be variable across agents. In our simulation, $r_c = 0.5$ for all agents.

Observed individual reputation is simulated in a similar way as encounter-derived reputation with the addition of observers. The setup mirrors observer-based image collection by Nowak and Sigmund (1998). Each agent a_i designates 10 random agents in the environment as being observed. All encounters by these 10 observed agents are recorded by a_i. The reputation of agent a_j in the eyes of a_i is r_{ij} which is the ratio of number of cooperation observed by a_i among its 10 observed agents' encounters over the number of defection. Such an RTFT agent a_i defects an opponent a_j if $r_{ij} < r_c$ where r_c is also set at 0.5 in the actual simulations.

Ideally, a_j's observed individual reputation should also depend on the mix of opponents that a_j has encountered. In our simulations, we are relying on the randomization process to reasonably sample the population for agents to be a_j's opponents. More sophisticated modeling would take the nature of a_j's opponents into consideration.

Group-derived reputation is simulated by grouping all agents with the same strategy into a group. The ***group reputation*** is calculated as the ratio of number of cooperation performed by members of a group over total number of encounters with a given agent. Reputation derived from a group depends on individual experience and is therefore not the same for all agents. When an RTFT agent meets an unknown agent, it uses the group reputation as the prior estimate for this unknown's reputation r_g. Such an RTFT agent defects an unknown opponent if $r_g < r_c$ where r_c is also set at 0.5 in the actual simulations. After the first encounter with an unknown agent, all subsequent decisions are based on encounter-derived individual reputation as discussed above.

Propagated reputation is simulated by having each RTFT agent recursively ask agents whom it has encountered before for their reputation estimate of an unknown agent. Propagation is checked by a MAX_TRAVERSAL limit. All gathered results are tallied using a Bayesian algorithm as described in the next chapter [2] for calculating the

[2] The algorithm is also described in Mui, *et al.* (2001).

propagated reputation r_p for an unknown opponent agent. If the calculated reputation $r_p < r_c$, the RTFT agent defects on the unknown opponent. Again, r_c is also set at 0.5 in the actual simulations. After the first encounter with an unknown agent, all subsequent decisions are based on encounter-derived individual reputation as discussed above.

6.1.6 Hypothesis

Our hypothesis is that reputation should lower the threshold of encounters per generation (EPG) necessary for TFT agents to dominate over AllD. By making TFT agents use different notions of reputation, we would like to compare how effectively each reputation notion allows the TFT agent to discriminate between AllD and other TFT agents in the environment.

6.2 Simulation Results

Figure 6.3 shows the evolution of TFT population size in a simulation starting with 50 AllD and 50 TFT agents. (No additional reputation measure is used by TFT agents except the 1 slot memory for the TFT for every one of its opponents.) The legend of Figure 6.3 indicates the number of encounters per generation (EPG). As the chance for repeated encounter is enhanced with increases in EPG, the TFT strategy dominates over AllD when EPG is greater than approximately 12000.

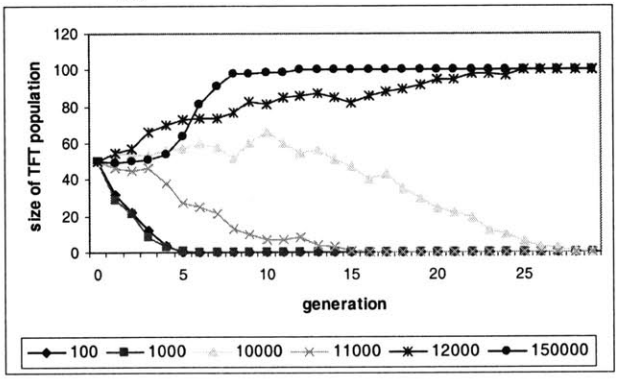


Figure 6.3 Base case when no reputation is introduced for TFT agents.

The same experiment as that shown in Figure 6.3 is done for each of the four notions of reputation as discussed in the last section. The EPG thresholds for RTFT strategies to dominate over AllD are summarized Figure 6.4.

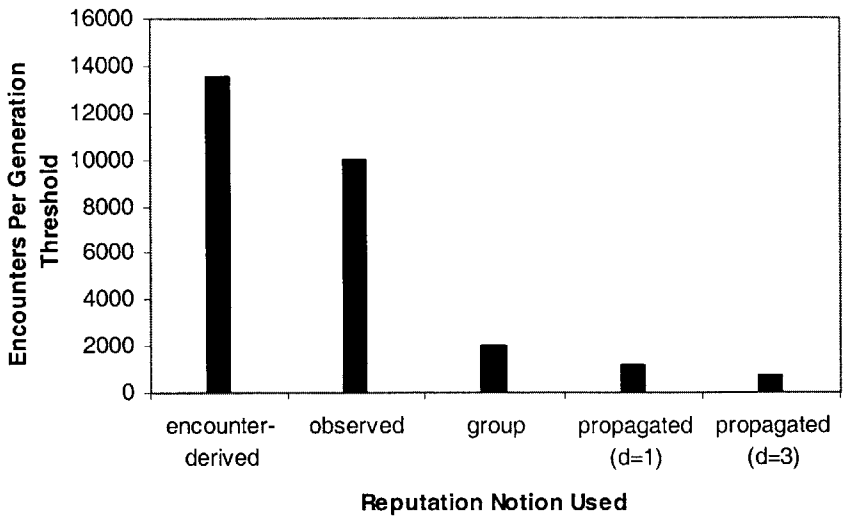


Figure 6.4. Number of encounters per generation (EPG) Thresholds for RTFT agents to become evolutionarily stable over AllD agents. 5 notions of reputation are used for RTFT agents, as indicated by the horizontal axis labels.

6.3 Discussions

Based on the encounters per generation (EPG) threshold, in order for RTFT agents to dominate over AllD agents, the following utility order is derived for the different notions of reputation in our simulations ('$\succ$' is the preference relation where $a \succ b$ indicates a is preferred over b):

$$r_{p3} \succ r_{p1} \succ r_g \succ r_o \succ r_e$$

Note the without using any aid from reputation information, the EPG for a population of TFT to dominate over AllD agents is 12000 as mentioned in Section 6.1. Therefore, an initial glance at Figure 6.4 might be surprising to find that strategies using encounter-derived reputation perform worse off then no reputation is used. One has to realize that in the evolutionary PD game that is simulated, encounter-derived individual reputation does not "kick-in" to warn an RTFT agent against an AllD agent until TFT agents have already cooperated once with an AllD agent. Therefore, the notion of direct encounter-derived reputation is not useful for this TFT-AllD game since repeated encounters between any two agents is rare.

Based on the size of drop in the number of encounters per generation (EPG) threshold, the propagated reputation seems to provide a significant utility to TFT agents against AllD agents. In other words, our proposed reputation framework in Chapter 5 has significant improvement over existing reputation schemes in terms of survival utility to agents in our simulated world.

We will return in the concluding chapter to a discussion of future works that will address whether the order of strength among the different notions of reputation holds in other types of game.

This chapter has described a simple simulation framework based on evolutionary games for understanding the relative strength of the different notions of reputation. Whereas these notions of reputation could only be compared qualitatively before, our simulation framework has enabled us to compare them quantitatively. The simulations show that (with the metric of encounters per generation as measure of strength), our proposed reputation notion performs very favorably in the evolutionary games compare to existing notions of reputation.

Our reputation simulation framework can be adapted to study an important problem for biologists: evolution of cooperation. In Chapter 7, we describe the intellectual space surrounding the study of evolution of cooperation. In Chapter 8, we will describe extension of our two-sided [3] Prisoner's Dilemma game from the earlier experiments in this chapter to a one-side game proposed by Nowak and Sigmund (1998) where the incentive for trust and reputation becomes more prominent.

[3] "Two-sided" refers to a payoff where both sides receive something in return (whether positive or negative) for every game that is played. "One-sided" refers to a game payoff where only one side is paid for every game.

CHAPTER 7

Evolution of Cooperation

Human have exhibited large scale cooperative behaviors that do not directly contribute to their individual survival (Fehr, *et al.*, 2000; Ensminger, 2002). Activities such as caring for the young and sick, altruistic act to the community, *etc.*, contribute to the public good. This raises the question for why people regularly engage in these cooperative activities – many of which are costly to the individual. We argue that the transmission of trust and reputation information is an important factor in the rise of cooperative behaviors in the evolution of a group of selfish individuals. This chapter provides the background for this argument. The main thrust of the argument is deferred to the next chapter.

Section 2.1 first provides the motivation and background for the study of evolution of cooperation. Section 2.2 reviews some game theoretic concepts that are frequently used in the study of evolution. Section 2.3 summarizes the strengths and weaknesses of major approaches that have been taken to study this subject. Section 2.4 concludes this chapter with thoughts on areas where this subject area can be further studied.

7.1 Motivation and Background

Darwin's theory of evolution argues for the "survival of the fittest" but does not specify the unit for fitness measure. By "fittest", one could mean the individual, the population, the gene, or the organization around a group of agents. Each interpretation calls for very different models for how natural selection works.

Rational agency theory underpins much of modern economics (Kreps, 1990; Samuelson and Nordhaus, 2000) and interprets "fittest" to refer to the individual. This individualistic view treats social organization as a by-product of self-interest. Phenomena such as altruism and cooperation are dismissed as no more than disguises with selfish motivations (Kreps and Wilson, 1982).[1] Such interpretations are in contradiction to sociological and biological observations of human and animal behaviors (Trivers, 1971; Ostrom, 1998; Fehr, *et al.*, 2000; Henrich, *et al.*, 2002). Reciprocative acts of giving and receiving permeate human life and the lives of many animals.

[1] Like Simon (1990), altruism is taken in this chapter to mean behaviors which on average increases the reproductive fitness of others at the expense of the fitness of the altruist. Fitness refers to the expected number of progeny. Cooperation often refers to altruistic acts where the cooperator puts future self-interest on the line for the recipient of these acts.

However, the individualistic view of evolutionary theory predicts that reciprocal altruistic cooperation should be limited to those who are "unfit" – natural selection should favor selfish behaviors.

The beginning half of the past century saw the proponents of interpreting "fittest" as that of the species or population (Haldane, 1932; Eshel, 1972; Wilson, 1980). This is known as the "group selection" theory. Although contested by later evolutionary theorists, group selection has for a time been able to account for why individual agents cooperate.

One dominant group of theorists interprets "fittest" to refer to the gene (Hamilton, 1964; Dawkins, 1979). Cooperation and related behaviors are explained in terms of fitness contribution to the individual gene. Individuals' aim in life is to maximize the chances that their set of genes (which also exists in others around them) is passed on to the next generation – even at the expense of the individual's fitness.

Yet another group of theorists considers "fittest" to refer to the social organization surrounding individual agents (Simon, 1969, 1990). This interpretation considers cooperation and sacrificial acts as enhancing the fitness of groups of agents within social structures, which in turn increases the individual fitness of its members.

This chapter summarizes five important approaches to the study of evolution of cooperation. Strengths and weaknesses of these approaches are discussed. Suggestions on how this field can move forward are provided.

	Group Selection	Kinship Theory	Direct Reciprocation	Indirect Reciprocation	Social Learning
Supporting	Eshel, 1972; Wilson, *et al.*, 1994;	Hamilton, 1963, 1964; Dawkins, 1976; Riolo, *et al.*, 2001;	Axelrod, *et al.*, 1981; Axelrod, 1984; Roberts, *et al.*, 1998;	Trivers, 1971; Alexander, 1987; Pollock, *et al.*, 1992; Nowak, *et al.*, 1998; Lotem, *et al.*, 1999; Wedekind, *et al.*, 2000;	Simon, 1969, 1991; Boyd, *et al.*, 1982;
Non-supporting	Fisher, 1958; Hamilton, 1963; Axelrod, *et al.*, 1981;	Murray, *et al.* 1984; Taylor, 1992; Wilson, 1992; Queller, 1994; West, *et al.*, 2001; Axelrod, *et al.*, 1981;	Alexander, 1987; Boyd, *et al.*, 1987; Nowak, *et al.*, 1998;	Boyd, *et al.*, 1989; Riolo, *et al.*, 2001;	N/A

Table 7.1. Representative works for five major approaches to the study of evolution of cooperation. Both supporting and non-supporting works for each approach are listed.

7.2 Iterated Games

Nearly all quantitative modeling in the study of evolution of cooperation uses game theory (Axelrod, 2000). The most common game studied for evolution is the prisoner's dilemma game. This section briefly reviews this game.

7.2.1 One Shot PD Games

This section describes in a bit more detail the Prisoner's Dilemma single stage game used in the reputation simulations in Chapter 6. For the PD game, the action space for each agent is:

- **Action** = { *cooperate*, *defect* }

The payoff matrix for the prisoners' dilemma game is given in Figure 6.1.

In the game-theoretic literature for describing payoffs for prisoners' dilemma (PD) game, recall from Section 6.1.2 the four descriptions are used for the outcomes of the game:

- *Temptation (T)*: when an agent defects (D) while its opponent cooperates (C)
- *Reward (R)*: when both an agent and its opponent cooperate (C) with each other
- *Punishment (P)*: when both an agent and its opponent defect (D) against each other
- *Sucker (S)*: when an agent cooperates (C) while its opponent defects (D)

With the following two constraints, the "dilemma" can be easily seen (Axelrod, 1984):

$$T > R > P > S$$

$$2R > T + S$$

The following assignments satisfy the above constraints:

$$T = 5, R = 3, P = 1, S = 0.$$

Given agent 2 is choosing to "cooperate," agent 1 can maximize its payoff by choosing "defect"; given agent 2 is choosing to "defect," agent 1 can maximize its payoff by choosing to "defect." Therefore, the *Nash Equilibrium* [2] of this one shot PD game is: Defect-Defect (the lower right quadrant of the payoff matrix).

By the rational agency theory of economics (Kreps, 1990), the utility maximization point is therefore *Defect-Defect*. However, both agents can do better by cooperating,[3] herein lies the dilemma.

[2] "Nash Equilibrium" refers to a utility-maximizing stability point where either agent cannot be better off by perturbing their actions at the equilibrium (Kreps, 1990).

[3] *Cooperate-Cooperate* is termed *pareto-superior* to the Nash Equilibrium point *Defect-Defect* in the one-shot PD game.

7.2.2 Iterated PD Games

If agents are indeed selfish utility-maximizing, economics tells us that in one shot PD games, they cannot do better than what is indicated by the Nash Equilibrium solution of defection. However, if the PD game is played multiple times between two agents, cooperation has some chance. In their influential 1980 *Science* paper, Axelrod and Hamilton introduced the "shadow of the future" (w) parameter and have shown how agents can develop cooperation when w exceeds a threshold in an iterated PD game. [4]

7.2.3 Evolutionary PD Games

To simulate the evolutionary process, birth and death of agents are introduced for multiple generations of iterated PD games among groups of agents. The most common scenario is the so called "non-overlapping" generations where all agents of one generation dies before the next generation is created (Axelrod and Hamilton 1984; Nowak and Sigmund, 1998). A generation is defined by a set of PD encounters among the agents. Participants in an encounter are chosen randomly from the total population. Often the goal of the evolutionary simulation is to assess the fitness of specific agent strategies or attributes.

In many studies, the goal is to understand the strategic implications for the survival of agents. To carry out these studies, at the end of each generation (where a certain number of dyadic encounters between agents have occurred), the total fitness of all agents with the same strategy is calculated. The number of progeny in the next generation with a given strategy is created in proportion to the total fitness associated with that strategy in the previous generation. The total population size is often fixed, so any increase in the number of one type of agent is balanced by a decrease in the numbers of other types of agents.

Commonly studied agent strategies are:

- All-defecting (ALLD): always defects.
- *Tit-for-tat (TFT)*: initially cooperates, and thereafter does what the other agent did in the last round.
- *Suspicious Tit-for-tat (STFT)*: initially defects, and thereafter does what the other agent did in the last round.
- *Grim-trigger*: initially cooperates, but once defected by another agent, this agent always defects against that other agent.

Axelrod (1984) has documented empirical evidences for the robustness of TFT compared to other strategies under competitive settings of many types of strategies.

[4] In the upcoming section on *direct reciprocation*, results from Axelrod and Hamilton (1981) will be discussed.

7.3 Existing Approaches to Study Evolution of Cooperation

This chapter examines five major approaches to the study of how and why evolution has evolved for social animals that exist today. Significant publications for each approach are briefly summarized. Table 1 highlights the representative publications for these approaches.

7.3.1 Group Selection

If the unit of natural selection is on the individual level, cooperation among individuals would not be consistent with Darwin's theory. A group of evolutionary theorists (Eshel, 1972; Wilson and Sober, 1994) therefore postulated the unit of natural selection is on the group level: species, community, population, *etc.*

Many argue that group selection theory is misguided (Axelrod and Hamilton, 1981). Difficulties with this theory arise from the strife often observed within species, communities, and populations. To explain for the manifest existence of cooperation and related behavior such as altruism, alternative theories are needed.

7.3.2 Kinship Theory

Kinship theory is based on the commonly observed cooperative behaviors such as altruism exhibited by parents toward their children, nepotism in human societies, *etc.* Such behaviors toward one's kin not only decrease individual fitness of the donor (while benefiting the fitness of others), they often incur costs – thereby decreasing personal fitness.

Hamilton's rule of relatedness provides the foundation of much of the work on kinship theory. This rule states that altruism (or less aggression) is favored when the following inequality holds:

$$rb - c > 0 \quad or \quad r > \frac{c}{b}$$

where r is the genetic relatedness of two interacting agents, b is the fitness benefit to the beneficiary, and c is the fitness cost to the altruist. This rule suggests that agents should show more altruism and less aggression toward closer kin. Hamilton (1964) uses Wright's Coefficient of Relatedness for r.[5]

Hamilton (1964) suggests that it is not the individual fitness that is evolutionarily selected but it is that of the "inclusive fitness" of a gene. *Inclusive fitness* measures the ability for a gene to reproduce itself in the offspring. This differs from the classical Darwinian notion of "individual fitness" in that the central actor of evolution is the hereditary unit and not the individual. Dawkins (1979) popularized this "selfish gene"

[5] Wright's Coefficient of Relatedness r=1 if the agent is compared to itself; r=0.5 for same parents siblings. r=0.25 for a grandparent and a grandchild; and so on. r can be thought of as the proportion of genes shared between two individuals.

thesis in his 1979 book. By using inclusive fitness for analysis, Hamilton and others are able to account for cooperative behaviors as means to maximize the inclusive fitness of individuals' genes.

7.3.2.1 Problems with Kinship Theory

A number of studies have contested Hamilton's thesis (Taylor, 1992; Wilson, 1992; Queller, 1994). The main difficulty pointed out is the interplay between relatedness and competition among kin. Factors that tend to increase the average relatedness of interacting agents, such as limited dispersal,[6] also tend to increase the amount of competition among relatives (Murray, *et al.*, 1984). West, *et al.*, (2001) have empirically examined fig wasp in nature and reported that with limited dispersal, the increased competition between relatives can negate the effect of increased relatedness in favoring altruism. Different fixes to Hamilton's rule have been proposed. West, *et al.*, (2001) modifies the rule by replacing b as follows:

$$b = B - a\,(B - c)$$

where c is the fitness cost as before and B is the benefit that would be given to the beneficiary in an interaction if competition does not exist. The parameter $a \in [0, 1]$ measures the extent to which neighbors compete (small a confers little competition while large a confers the opposite). $a = 0$ equals the original Hamilton' rule.

Kinship theory also has difficulty in explaining cooperation where relatedness is low or absent. Axelrod and Hamilton (1981) pointed out examples from mutualistic symbioses such as that between fungus and alga that compose a lichen; the fig wasps and fig trees where wasps serve as the tree's sole means of pollination and seed set. Furthermore, cooperation in such symbioses can sometimes turn into antagonism (Caullery, 1952). Kinship theory cannot explain this dynamics of cooperation at all. The theory of reciprocation answers these critiques.

7.3.3 Reciprocation Theory

Trivers (1971) postulated "reciprocal altruism" as an explanation for how individuals are willing to sacrifice personal gain for the good of another. Specifically, phenomena such as friendship, moralistic aggression, gratitude, sympathy, trustworthiness, *etc.* can be explained in terms of deferring immediate personal gain toward potential benefits from future reciprocations by others.

Much of the work on reciprocation theory is carried out using a game theoretic framework based on the Prisoner's Dilemma (PD) game and the Tit-for-Tat (TFT) strategy, which have been discussed earlier in this chapter.

[6] Dispersal refers to the geographical distribution that a group of agents are placed. This often refers to individuals who are related, as in kin selection theory.

7.3.3.1 Direct Reciprocation

The strategy TFT has been postulated by Axelrod and Hamilton (1981) to be an *evolutionary stable strategy* for the iterated PD game (Maynard Smith, *et al.*, 1973),[7] provided with a sufficiently large probability of repeated future encounters between agents. They label this probability "shadow of the future" (w). In other words, cooperation between agents can evolve using the reciprocating strategy of TFT.

TFT is a *direct* reciprocating strategy for agents in the PD game – reciprocating what the other agent did in the previous round. In the context of an iterated PD game with only TFT agents and ALLD agents (*c.f.*, Section **Error! Reference source not found.**), the threshold w for persistent cooperation to develop can be calculated analytically. Consider a game with n TFT agents against n ALLD agents, a given w, and the PD payoff matrix shown in the Appendix:

Payoff for a TFT agent against a TFT agent:

$$R\,(1 + w + w^2 + \ldots) = R/(1-w)$$

Payoff for a TFT agent against an ALLD agent:

$$T + P(w + w^2 + \ldots) = T + wP/(1-w)$$

For TFT agents to defend themselves against ALLD agents, the following inequality must hold:

$$R/(1-w) > T + wP/(1-w) \qquad or \qquad w > (T-R)/(T-P)$$

Experimental findings of TFT behaviors have been found for tree swallows (Lombardo, 1985) and sticklebacks (Milinski, 1987) in nature.

Other theoreticians have disapproved of TFT's significance as a workable strategy toward cooperation. Specifically, Boyd and Lorberbaum (1987) have shown that there does not exist a single strategy (including TFT) which can be resistant to all kinds of invading strategies for the PD game. The counter proof they use to show that TFT is not an evolutionarily stable strategy is by introducing two others: the Tit-for-two-tat (TFTT) and Suspicious-tit-for-tat (STFT) strategies. A TFTT agent does not retaliate until there have been two successive defections; a STFT agent starts off defecting on the first encounter but thereafter plays TFT.[8]

7.3.3.2 Indirect Reciprocation

It is a common practice in human societies where the donor of a good deed might not necessarily be rewarded by the recipient directly but by other individuals who might be recipients of other good deeds by others. Alexander (1987) is the first to term this phenomenon "indirect reciprocity."

---

[7] "Evolutionarily stable strategy" or ESS is a term coined by Maynard Smith and Price (1973) to refer to a strategy such that if most members of a population adopt this strategy, there exists no "mutant" strategy that would give high reproductive success.

[8] If a group of agents consists of some who use TFT, STFT, or TFTT strategies, those using TFTT have the edge over the other two. Therefore, TFT is not evolutionarily stable. Carefully syncing TFTT's moves with alternating "cooperate" and "defect" can invade a group of TFTT easily. Hence, neither is TFTT evolutionarily stable.

Boyd and Richerson (1989) are the first to develop a mathematical model of cooperation based on indirect reciprocity. In their model, donors are rewarded for their deeds by the last person in a ring of *n* indirectly-reciprocating individuals. Their analysis of the indirect reciprocity indicates that the conditions necessary for the evolution of indirect reciprocity "... become restrictive as group size increases." Their analysis however does not take into account social structures other than ringed interactions. Human agents and their interactions tend to form embedded graphs which are not rings (Granovetta, 1985).

Nowak and Sigmund (1998) developed an indirect reciprocation model based on image (reputation) scoring. In their non-overlapping evolutionary model, each agent has a genetic strategy and a non-heritable image score.[9] They showed that cooperation can be established under "global" image scoring if the number of interactions per generation is sufficiently large. "Global" refers to the nature of an agent's image being visible to all other agents. Following the results by Pollock and Dugatkin (1989), Nowak and Sigmund (1998) further studied the effect of having randomly selected observers on the evolution of cooperation. Every interaction is no longer globally known but is observed by a limited number of observers. They conclude that cooperation may evolve through indirect reciprocity with or without global knowledge about agents' image scores. Wedekind and Milinski (2000) have experimentally verified Nowak and Sigmund's hypothesis that image scoring does play a role in actual human cooperation.

Extensions to Nowak and Sigmund's work have been done in two areas. Lotem, *et al.* (1999) have included persistent non-cooperators ("phenotypic defectors") to model after the sick, young or handicapped who may be unable to cooperate even if they are genetically predisposed to do so. Their simulations show that phenotypic defectors paradoxically allow persistent discriminating cooperation under a much wider range of conditions than found by Nowak and Sigmund.

Riolo, *et al.* (2001) have infused kinship theory into Nowak and Sigmund's model with "tag-based" reciprocity – cooperation based on inheritable and identifiable tags on agents. Their simulations indicate that cooperation can evolve even when reciprocity is absent.

7.3.4 Social Learning

Boyd and Richerson (1982) introduce a model for cooperation to evolve based on "conformist transmission," or *cultural transmission.* This mechanism refers to the preferential selection of the behaviors individuals encounter most frequently. In other words, individuals learn the most dominant behaviors in their embedded social network. Although cooperation can evolve using this mechanism, conformers and non-conformers are not distinguished. Along this line, Simon (1990) proposes an alternate social learning approach.

Simon (1990) argues that in human societies, cooperation exists where kinship and reciprocation are absent. Specifically, he uses the case of altruism often documented in

[9] Non-overlapping refers to a common simulation simplification that no 2 generations of agents exists in the same time. Refer to the Appendix on evolutionary PD games for further explanation.

organizational studies where real people who are satisfied with their positions adopt the organization's interest as their own interest.

Simon (1990) considers a population of two types of individuals in his model: A and S, in proportions p and $1 - p$, respectively. Population size is n. Individuals of type A are altruistic (who always cooperate) while those of type S are selfish. Each A exhibits behaviors that contributes b offspring to members of the population (including himself). The cost to A is c fewer children for A. The average number of offspring, F_A and F_S, of each A and S in the absence of social learning is:

$$FA = X - c + bp$$

$$FS = X + bp$$

where X is the number of offspring in the absence of altruistic behavior. To clarify, any individuals can be recipients to the np altruists, and selfish S individuals have no altruism costs. Without social learning, since $c > 0$, altruists always have fewer children than selfish ones. In evolutionary game theory terms, selfish individuals are more fit and will therefore have the evolutionarily stable strategy over altruistic ones.

To model social learning (learning from others in the society), Simon introduces the *docility* parameter (d), which refers to the willingness of an individual to " ... accept well the instruction society provides them." The content of what is learned will not be fully screened for its contribution to personal fitness. Docility models the level of *bounded rationality* and computational limitation of individuals. The higher the docility, the more boundedly rational and the more computationally limited an individual is. F_A and F_S can be modified as follows:

$$FA = X + d - c + b(c)\,p$$

$$FS = X + b(c)\,p$$

b is now a function of c because the amount of altruism exacted from A depend on the society's definition of proper behavior. The condition for cooperation (and altruists) to dominate is:

$$d - c > 0$$

No experimental work verifying or disapproving Simon's model is known to the author.

7.4 Extending Existing Models

Research attempting to explain how cooperation arises in society has been at least a century old. Relative to many other research problems, this is a mature field with the wisdom of many very intelligent and dedicated researchers. Nonetheless, existing models fall far short of fully explaining how and why cooperation evolves in natural population.

7.4.1 Unifying Perspectives

Firstly, existing models and theories appear to have conflicting assumptions and implications, as illustrated in the two rows of supporting and non-supporting publications in Table 1. Each of the five major approaches discussed in this paper seems to explain certain aspects of cooperation and how it evolves but is lacking in other aspects. Clearly, cooperation does evolve. Therefore, a unified theory of how and why cooperation evolves still eludes the research community.

7.4.2 Towards Realistic Models

Secondly, most models are still based on simple games – as as prisoner's dilemma, which are a gross over-simplifications of actual agent-to-agent interactions. Important elements comprising actual animal interactions are missing in existing models. Glaringly amiss are the following:

- Nearly all existing models are based on asexual, non-overlapping generations of agents. Obviously, these are not realistic assumptions.
- Simultaneous play for every interaction is assumed in all but a few models (e.g., Abell and Reyniers, 2000)
- Interactions are restricted to dyadic. Clearly, animals are often involved in interactions involving more than one other animal. Studies into other types of interactions are much needed.
- Agent behavior is assumed to be unmistakable. As pointed out by May (1987), human behavior is far from mistake-free. Future studies of how cooperation evolves should take the stochasticity of behavior seriously.
- Actions are discrete in existing models. Clearly cooperation is not just "cooperate" and "defect" alone but the levels of cooperation are continuous. Perhaps this continuity is determined by the amount trust between agents.
- Animals are not born with specific and unyielding strategies. Depending on the situation, animals might switch between different types of strategies in cooperating with others. Strategy modification and defection should be studied.
- Although works by Simon (1990, 1991), Cohen, et al., (2001), and others have brought social structure and learning into the study of cooperation, much work is still needed to understand how and to what extent various types of social structure affect cooperation. Given the importance placed on cooperation across social and

economic organizations, this should be of enormous interest not only to the research community.

It is hard to believe given the century old nature of studies on evolution of cooperation, there exist so many glaringly unrealistic aspects in current models. This lack of progress points to the difficulty and mysteriousness of this natural phenomenon. It also points to the research potential of this field.

Throughout evolution, humans have exhibited cooperative behaviors which do not directly contribute to their individual survival. Activities such as warfare, caring for the young and sick, altruistic acts to the community, *etc.* contribute to the public good. This raises the question: why people regularly engage in these cooperative activities – many of which are costly to the individual. This chapter has summarized the major approaches that have been taken to answer this question. After over a century of research, the answer to this question remains open. This field continues to provide a fruitful area for evolutionary research. The next chapter presents one such fruitful approach based on "social information."

CHAPTER 8

Evolution of Cooperation by Social Information

As reviewed in the last chapter, the complexity of humans' cooperative behavior cannot be fully explained by theories of kin selection (Hamilton, 1963, 1964) and group selection (Williams, 1971; Wilson and Sober, 1994). Reciprocity approaches hold promising results. Among the two main reciprocity approaches – direct and indirect, direct reciprocity is limited by the size of the population because in large populations, most agents do not interact with most others.

If indirect reciprocity – or "reciprocal altruism" (Trivers, 1971; Axelrod and Hamilton, 1981; Axelrod, 1984; Axelrod and Dion, 1988) – is to provide an explanation for altruistic behavior, it would have to depart from direct reciprocity, which requires dyads of individuals to interact repeatedly. For indirect reciprocity (Alexander, 1987; Boyd and Richerson, 1989; Pollock and Dugatkin, 1992) to rationalize cooperation among genetically unrelated or even culturally dissimilar individuals, information about the reputation of individuals must be assessed and propagated in a population.

In this chapter, we propose to apply our rating propagation schemes from Chapter 3 and 5 to the problem of evolution of indirect reciprocity. We term the information gathered from direct and indirect inference "social information": information retrieved from and propagated through dynamically evolving networks of trust and reputation. We argue that cooperation is an act of trust and is sustained by reciprocity and propagation of reputation information in a social environment. We detail the computational model of this assertion in this chapter.

We refer to the information content of such networks of trust and reputation as the 'collective memory'. We show that for indirect reciprocity to be evolutionarily stable, the ratio of the probability of trusting and helping a reputable individual to the probability of helping a disreputable individual must exceed the cost-to-benefit ratio of the altruistic act. In other words, the benefit received by trusting the trustworthy must out-weigh the cost of helping the untrustworthy.

8.1 Social Information: Role in Evolution of Cooperation

Alexander (1987) coined the term *indirect reciprocity* to refer to the commonly practiced act of reciprocation in human societies, where the donor of a good deed does not necessarily expect to be rewarded by the recipient but perhaps by another individual who may be the recipient of other good deeds by other donors. He has hypothesized that "indirect reciprocity is a consequence of direct reciprocity occurring in the presence of others" (*ibid.*). For indirect reciprocity to work, Alexander conjectures that reputation and status of members in a group must be continually assessed and reassessed. Hence, reputation is a key concept in Alexander's premise.

Several authors have since attempted to formalize the mechanisms by which indirect reciprocity can evolve. Boyd and Richerson (1989) developed a mathematical model of 'circular reciprocity' where the donor of a good deed is to be rewarded by the last individual in a ring of n reciprocating individuals. Their results suggested that indirect reciprocity is unlikely to be important unless interacting groups are relatively small.

Alexander (1987) has further hypothesized that "indirect reciprocity is a consequence of direct reciprocity occurring in the presence of others." Those who observe direct reciprocation between individuals will then be in the position of assessing the *reputation* of members of a population. Hence, reputation is a key concept in Alexander's premise. Under this premise, indirect reciprocity requires that the reputation and status of members of a group be continually assessed and reassessed.

Pollock and Dugatkin (1992) investigated the significance of observation in guiding behavioral choice by studying a variant of tit-for-tat (TFT) where players behave like TFT in the absence of information about a new co-player but defect if the co-player has been observed defecting in his last interaction. The result of this investigation is that when TFT fails to be *evolutionarily stable*,[1] whereas the variant using observers is. They term the information used to aid selection of action *reputation*.

Following Alexander's conjecture and studies by Pollock and Dugatkin, Nowak and Sigmund (1998, 2000) developed a model of indirect reciprocity by image scoring to study the role of observers in assessing the reputation of members of a population and eventually on the evolution of cooperation. Under this model, every player (when selected as donor) has an image score which is modified locally only by the recipient of the action and a few randomly selected observers. Hence, different individuals may have different perceptions about the same player. In the terminology of Chapter 3 and 4, different individuals have different "ratings" about the same player.

Cooperators are rewarded for their altruistic acts through increases of their image scores for the game recipients and observers. Every player also has a strategy value.[2] When playing against a recipient, the donor compares the image score of the recipient to his own strategy and cooperates only if the recipient's image is at least as high as his own strategy k. The underlying premise in this framework is that if the information about the

[1] *Evolutionary stability* is an equilibrium condition first defined by Maynard-Smith and Price (1973) to describe conditions under which agents with certain strategies can dominate a population.

[2] The strategy value is an integer k and $-5 \leq k \leq 5$ such that the agent only donates to a recipient if the recipient's image is greater than or equal to k.

image of members of a population can be obtained, then an informed donor would only helps those who are likely to help back in the future. This in turn will improve the reputation of the donor. This strategy introduces feedback into the system so that although an altruistic act of cooperation entails a cost, as a donor's reputation is enhanced, the likelihood for others to help (*i.e.*, increase the fitness of) this donor in the future is increased.

The underlying premise in this framework is that if information about the image of members of a population can be obtained, then an informed donor will be able to help those who are likely to help back in the future. To examine the role of observers as the source of information about players' actions, Nowak and Sigmund simulated an environment where every round of a game was observed by 10 randomly selected observers. These observers plus the recipient are the only ones who can update their perception of the donor's image. Hence, different individuals may have different perceptions about the same player. For this information to be of utility, observers and recipients have to interact with observed agents. Their results suggested that when information in localized, cooperation can still be established in populations, but a greater level of interactions is needed for cooperation to be sustainable in larger populations. Their work well demonstrated the significance of information about the reputation of interacting individuals in the evolution and sustenance of indirect reciprocity.

8.1.1 Trust and Reputation in Social Networks

Although making observations is one of the mechanisms for acquiring information, it is not the principal manner by which humans process information about others' actions. Furthermore, reputation is seldom propagated in a random manner in populations. Embedded in every social network is a web of trust with nodes representing members of the web and edges representing the amount of trust among pairs of acquaintances. When faced with social dilemmas, such as to cooperate or not, individuals make use of social information available to them to reduce uncertainty.

A key mechanism by which humans acquire information about others is by seeking the opinions of trusted and reputable acquaintances. Parents are the first members of such networks of trust who provide their children with instruction and advice. Because it is not possible for one to observe or remember all possible events even in small populations, the information content of such a web of trust serves as one's *collective memory*. Such trust-based body of information seems to be a fundamental element of social information and an inherent aspect of the processes by which humans make decisions.

Aside from this "direct" notion of social information, where it is assumed that all members of a social network are acquaintances, there is another vital, while not as apparent, component of social information, where one attempts to seek the opinion of a reputable 'k-degree acquaintance' who is not a direct acquaintance but is connected through a chain of k other individuals. In other words, trust can be inferred in a transitive manner using the notion of reputation (Mui and Mohtashemi, 2002).[3] Such a process of

[3] An example of such transitivity is: if agent a trusts b, who in turn trusts c. Even if a has not met c before, if c's good reputation is communicated to a by b and perhaps other sources, a can be understandably cooperate with c in their first interaction.

decision making not only requires the ability to induce friendship and make acquaintances, it also entails the cognitive ability to reason, learn, and communicate. One can speculate that the evolution of language and intelligence plays a critical role in the evolution and sustenance of indirect reciprocity. What seems common in most models of evolution of cooperation is the players' passive treatment of information and their limited ability to remember only the near past. Even if one assumes, to keep matters simple, that human memory is not functional on a longer time line, it is important to note the existence of a collective memory on which humans routinely rely on to gain more information. However, some means of communication is required for processing and retrieving the information content of the collective memory.

8.1.2 Dynamics of Social Networks

The social interactions discussed thus far assume dynamicity in the underlying social networks. Even if one assumes closed populations of constant size, due to mobility and communication, new links are created and new acquaintances are made at all times.[4] This means that the topology of social networks change over time. As new links are added to the network the path between every two individuals become shorter over time.[5] In other words, the probability that any two randomly selected individuals know each other increases with time, thereby modifying the content of the collective memory in a dynamic manner.

In the next section, we will present a simulation framework which has the following assumptions:

- Players can communicate and inquire about the reputation of their co-players.
- The networks of acquaintance grow dynamically over time.
- Information about the reputation of co-players is not obtained randomly; rather, players selectively acquire information from their acquaintance network by taking advantage of the collective memory of the social networks to which they belong.
- Information is not propagated randomly in a population. New information resulting from new interactions modifies the content of the collective memory of a recipient and is therefore selectively propagated through the recipient's acquaintance network.
- A social network is an evolving dynamic entity. With new interactions, new links are created which in turn increase the likelihood of any two randomly selected players would *know* each other.

---

[4] The game theoretic model to be simulated uses so called non-overlapping generations: no one dies before the end of each generation. Hence, our model here assumes that links are only added and not removed.

[5] A path between two individuals in a social network is defined as the number of distinct intermediary nodes connecting them. Geodesic distance in a graph is a good example (Wasserman and Faust, 1994).

8.2 Social Information : Simulation Framework

To make concepts related to social information clear, we propose to simulate a simple model of the above framework. We simulate a series of evolutionary games whose stage game can be considered as a "one-sided" Prisoner's dilemma where only one player's payoff is affected at every stage. The coupling among the players' payoff is achieved through a parallel game involving adjusting the "image" (or reputation) values of the players.

Consider a population of n individuals divided into non-overlapping groups of acquaintances of the same initial size, s:

n : population size

s : acquaintance size for every individual at the beginning of each generation

Here, we model the underlying graph structure of a group as a clique. (In the remainder of this chapter we will use the terms group and clique interchangeably.) Every generation consists of a fixed number of rounds:

m : number of rounds per generation

In every round two players are selected at random: one as donor and the other as recipient. The donor has an option of cooperating with or defecting upon the recipient. If the donor cooperates it will cost him [6] a value of c and the recipient receives a benefit value of b ($b>c$):

b : benefit *per round for the recipient who receives a donation*

c : *cost per round* to a donor who gives a donation

If the donor fails to cooperate, no one gains any benefit nor incurs any cost. However, the donor's image will suffer, as discussed below.

At the beginning of each generation every player is born into a unique clique of acquaintances. At the end of each generation everyone dies and produces offspring in proportion to the total payoff they receive throughout their generation. Modeling after the simulation framework by Nowak and Sigmund (1998), every agent j possesses a strategy, k_j, and an image score about agent i, s_{ij}:

k_j: agent j*'s strategy (cooperate if the recipient's image* $\geq k$*)*

s_{ij}: *image that* agent j has about agent i

At the beginning of a generation all players have image score of zero. A potential donor cooperates if the image score of the recipient is at least as high as his own strategy value (k). The image scores of players are only known to and updated for their acquaintances. If a potential donor cooperates, his image score is increased by one unit; otherwise it is decreased by one unit. The notion of acquaintance set is treated dynamically here. A donor performs one of the two actions, cooperate or defect.

Donor's action space: { cooperate, defect }

[6] We use male pronoun for donor and female pronoun for recipient.

After a donor's action, the recipient and members of the donor's clique update their perception of the donor's image; the donor will then become a 'one-way acquaintance' of the recipient, *i.e.*, if in future rounds the recipient is chosen as a donor, in addition to her acquaintances, she will also ask the donor about her opponent. In other words, the donor is being added to the recipient's acquaintance set but not vice versa.

This asymmetric setup for whether the donor or recipient is added to the acquaintance set of the other was originally meant to represent the asymmetry in many everyday interactions. For example, the sellers of goods are more likely to know more about the goods than the buyers do. In our simulations, the donor gains no information about the recipient by donating to the recipient, but the recipient who has observed the donor in action can use this information for future encounters.[7]

If a potential donor j does not know the image score of the recipient i, he will make use of the social information available by asking all his acquaintances whether they have ever played in recipient role against the current recipient, *i.e.*, if the current recipient is a one-way acquaintance of any one in the donor's acquaintance set $\boldsymbol{Q}_j$. If no information is learned, the donor will assume an image score of zero; otherwise, he assesses the reputation of the recipient by adding up her image scores, provided by members of the acquaintance set $\boldsymbol{Q}_j$, and dividing by the total number of encounters.

Reputation of a recipient i in $\boldsymbol{Q}_j$: ratio of cooperation over all j's encounters in $\boldsymbol{Q}_j$

Such an averaging scheme has the benefit of transparency during analysis.

A potential donor j compares the computed reputation score s_{ij} for a potential recipient i to his own strategy k_j. In Nowak and Sigmund's scheme, j donates if $s_{ij} \geq k_j$. Therefore, the outcome of a round depends on the probability of knowing the recipient's image, which is derived from the collective memory embedded in one's acquaintance set. Furthermore, the underlying social structure is itself evolving as players meet over time, which causes the probability of knowing a randomly selected recipient to increase over the lifetime of a generation. The next section analytically describes the dynamics of this probability.

8.3 Social Information : Analysis

For the analysis of the framework presented in the past section, several variables need to be defined:

A_i : the *average* number of acquaintances per player at round i

This variable represents the average network connectivity per player at round i. Then A_0 is the initial clique size at the beginning of each generation (a clique includes self and acquaintances).

[7] The details on the authenticity of our modeling have unfortunately escaped our attention; and we cannot claim that our asymmetric setup mirrors real world interactions. In particular for our model, any agent in recipient's clique can consider the donor and members of the donor's clique as acquaintances. The asymmetry lies in that the donor and members of the donor's clique do not consider the recipient nor the recipient's clique as acquaintances.

q_i : the probability that a potential donor knows the image score of the recipient at round i.

In terms of A_i, q_i can be expressed as:

$$q_i = \frac{A_{i-1} - 1}{n-1} \tag{1}$$

i.e., the likelihood of knowing the recipient at round i is one less than the average number of acquaintances per player from the previous round (assuming that a player cannot play as both recipient and donor) over n-1 possible acquaintances one can have in a population of size n. The average network connectivity per agent at round i can be expressed as the following recurrence:

$$A_i = A_{i-1} + (1 - q_i)\frac{A_{i-1}}{n} \tag{2}$$

This relation shows that a new link is created between two players in every round of the game only if the donor does not know the image score of the recipient, in which case the donor is added to the acquaintance set of the recipient resulting in as many as A_{i-1} new links.

Rewrite A_i using equation (1), the following can be derived:

$$A_i = (1 + \frac{1}{n-1})A_{i-1} - \frac{{A_{i-1}}^2}{n(n-1)} \tag{3}$$

Substituting $X_i = A_i / n^2$, equation (3) can be expressed as the following canonical form:

$$X_i = (1 + \frac{1}{n})X_{i-1}(1 - X_{i-1}) \tag{4}$$

which is the familiar logistic equation.

As $n \to \infty$, the growth rate, $\left(1 + \frac{1}{n}\right) \to 1$. Therefore, for $n \geq 2$ the system is stable with the nontrivial fixed point $A^* = n$. This is the maximum network connectivity per player, *i.e.*, the maximum number of acquaintances an individual can have in a population of size n (including self).

If on the other hand, we assume that in each round all players are paired up to play as either recipients or donors, as in the remainder of this section, then the definition of q_i must be slightly modified to represent the probability that in each round, for each pair, a potential donor knows the image score of his co-player. A donor, on average, has $\frac{A_{i-1}}{2}$ acquaintances that can be selected to play as recipient. This is because, on average, half of the donor's acquaintances will be chosen as recipients and half as donors. There are

$n/2$ donor-recipient pairs (for n even), which means there are $n/2$ players out of which the recipient can be selected. Therefore, assuming that in each round everyone plays as either recipient or donor, for each pair, the probability that the donor knows his opponent will be: $q_i = \frac{A_{i-1}/2}{n/2} = \frac{A_{i-1}}{n}$. This variable, regardless of whether we assume one donor-recipient pair or $n/2$ donor-recipient pairs per round, amounts to the average information about the reputation of a co-player, retrievable from the collective memory embedded in the acquaintance set of a discriminating donor at a point in time.

Now consider the simulation framework for incomplete information by Nowak and Sigmund (1998, 2000), where a population of size n consists of two types of players: defectors who never help and discriminators who only help players with good image. Let the frequencies of these populations be:

x : the frequency of discriminators

y : the frequency of unconditional defectors

For a discriminating donor, a recipient has a good image, G, if she is known to have cooperated the last time; otherwise she has a bad image, B. We modify this model by imposing a *social structure on acquaintance relations*. If discriminating donors do not learn new information about their recipients by asking their acquaintances, they always cooperate. This means that in the absence of information, defectors can be mistaken for good scorers by discriminators. Therefore at the beginning of each generation, when no information about players' behaviors is available, all players are assumed to be good scorers. Let $A_x(i)$ and $A_y(i)$ denote the respective payoff for discriminators and defectors at round i. These two quantities can be expressed as:

$$A_y(i) = \frac{b}{2}\{x(1-q_i) + xq_i \frac{y_G(i)}{y}\}$$

$$A_x(i) = \frac{1}{2}\{-c(q_i g_i + 1 - q_i) + bx - bxq_i(1 - \frac{x_G(i)}{x}\}$$

where $x_G(i)$ and $y_G(i)$ are the respective frequency of good scoring discriminators and defectors at round i. These frequencies can be expressed as:

$$x_G(i) = \frac{1}{2}\{x_G(i-1) + x(1 - q_{i-1} + q_{i-1}g_{i-1})\}$$

$$y_G(i) = y2^{1-i}$$

The differential payoff at round i for discriminators is:

$$D_x = \frac{1}{2}\{-c(q_i g_i + 1 - q_i) + bq_i(g_i - 2^{1-i})\}.$$

Here $g_i = x_G(i) + y_G(i)$ is the proportion of good scorers at round i. Because everyone is assumed to have a good image at the beginning of each generation:

$$g_i = \begin{cases} 1 & i = 1 \\ \frac{1}{2}\{g_{i-1}(1+q_{i-1}x)+(1-q_{i-1})x\} & 2 \le i \le m \end{cases}$$

The differential total expected payoff for discriminators is then:

$$P_x = \frac{1}{2}\{-c[m-\sum_{i=1}^{m} q_i(1-g_i)]+b\sum_{i=1}^{m} q_i(g_i-2^{1-i})\} \tag{3}$$

For $P_x > 0$, we must have:

$$c\{m-\sum_{i=1}^{m} q_i(1-g_i)\} < b\sum_{i=1}^{m} q_i(g_i-2^{1-i}) \tag{4}$$

Divide both sides of inequality (4) by m to get:

$$c\{1-\sum_{i=1}^{m} q_i(1-g_i)/m\} < b\sum_{i=1}^{m} q_i(g_i-2^{1-i})/m \tag{5}$$

To interpret inequality (5) in an intuitive manner note that the following ratio is the average per round probability of knowing and helping a good scoring discriminator, *i.e.*, the probability of a discriminator making the right decision:

$$\sum_{i=1}^{m} q_i(g_i-2^{1-i})/m$$

Similarly, the following ratio is the average per round probability of knowing and not helping a bad scorer:

$$\sum_{i=1}^{m} q_i(1-g_i)/m$$

Hence, the term $1-\sum_{i=1}^{m} q_i(1-g_i)/m$ describes the average per round probability of knowing and helping a bad scorer, *i.e.*, the probability of a discriminator making the wrong decision. Inequality (5) then asserts that for discriminators to outperform the defectors the average per round benefit received by a good scoring discriminator must out-weigh the average per round cost to a discriminator for helping a bad scorer. Therefore trust pays off if it is based on information and placed upon the trustworthy.

If we rewrite inequality (5) in terms of cost-to-benefit ratio we derive that for discriminators to be evolutionarily stable we must have:

$$c/b < \frac{\sum_{i=1}^{m} q_i (g_i - 2^{1-i})/m}{1 - \sum_{i=1}^{m} q_i (1-g_i)/m} \qquad (6)$$

this is the ratio of knowing and helping a good scoring discriminator to knowing and helping a bad scorer must out-weigh the cost-to-benefit ratio. This result should be compared to the result from (Nowak and Sigmund, 1998). Under similar assumptions on the population composition but without a dynamic social structure, Nowak and Sigmund derived the following interesting condition for discriminators to be evolutionarily stable:

$$q > \frac{c}{b}$$

where q is a constant representing the probability of knowing the co-player's image score. Nowak and Sigmund aptly pointed out the similarity between their result and Hamilton's rule for altruism through kin selection (Hamilton, 1964), where the parameter for genetic relatedness was replaced by q, *i.e.*, the likelihood of knowing the opponent's reputation. However, when we impose a constantly evolving social structure for trust and acquaintanceship into the same framework, the parameter of 'familiarity', whether genetic, cultural, or simply due to observation or interaction, is replaced by a variable that amounts to the likelihood of knowing and trusting the reputation of the opponent.

A property of the framework outlined here is that the underlying network of trust relations is itself evolving as players interact. Hence, A_i and q_i are changing over time. As the game continues therefore, the threshold on c/b increases (see inequality (6)). Under such changing environment the payoff to cooperative strategies may be negative at the beginning, but over time they may be able to recover and even outperform the exploiters. This is achieved through two parameters in the system: the initial clique size, A_0, and the number of rounds, m. The effect of varying the initial clique size on the long term outcome of the game will be demonstrated in a set of simulations in the next section. But even under fixed initial clique size, the probability of knowing the image of an opponent is increased over time – following the equations (2)-(4). On the other hand, a slight increase in the number of rounds can help cooperation be established even in a population of 'all-loners'.

8.4 Social Information : Simulation Results

Simulation results in this section are based on the experiments as described in Section 3.2 and as analyzed in Section 3.3.

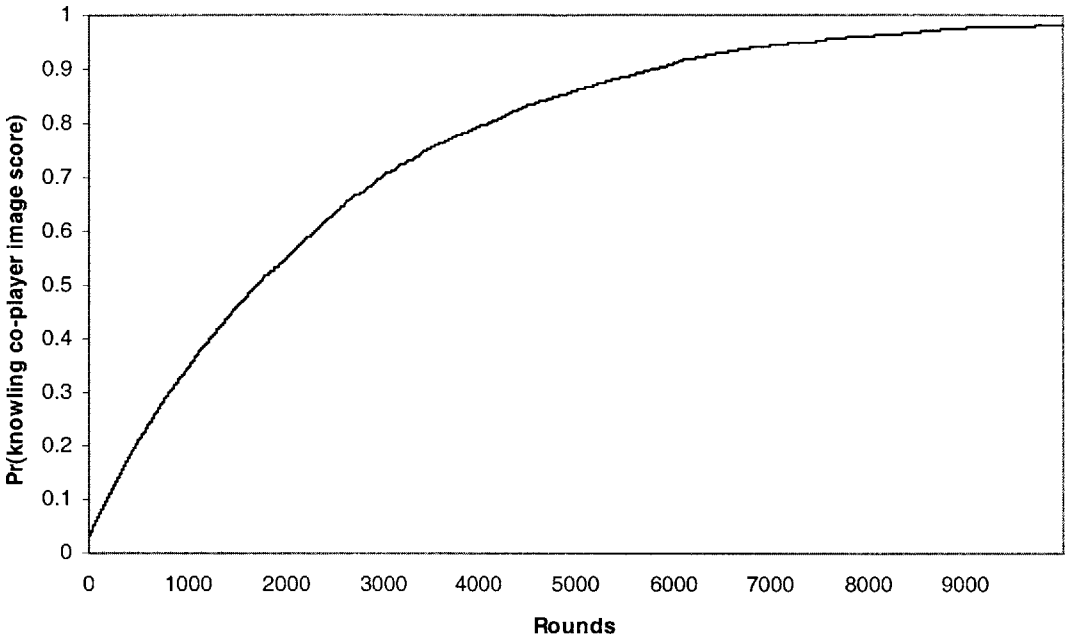


Figure 8.1. Dynamics of acquaintanceship.

Figure 8.1. shows that the probability of knowing a randomly selected co-player's image score increases with number of rounds. Initially a population of n=100 individuals is divided into non-overlapping cliques of size 4. As the game is continued, agents meet and make new connections. The average number of acquaintances per player increases with every round of the game, and therefore the probability of knowing the opponent's image also increases. Here a generation consists of 10,000 number of rounds. The rise in the probability of familiarity is consistent with the analytic result in Equation (2) in the former section.

Figure 8.2 shows the results of computer simulations for a population of n individuals with an initial acquaintance clique of size four. Modeling after Nowak and Sigmund (1998)'s simulation framework, the strategy k ranges from -5 to 6 where k=-5 represents unconditional cooperators, k=6 represents defectors, and k=0 represents the most discriminating. The image scores range from -5 to 5. A potential donor cooperates only if the image score of the recipient is at least as large as his own strategy. The children inherit the strategies of their parents unless they are subject to mutation at a rate of 0.001. We sampled the frequency distribution of strategies over 106 generations for population sizes n=52, n=100, and n=200. Every generation consists of a fixed number of rounds, $m=10n$. All other parameters are as in Nowak and Sigmund (1998)'s experiments.

Under these assumptions, the likelihood for a pair of individuals to meet more than once is negligible. Children inherit neither the image score of their parents nor their parental acquaintance structure. They only inherit the strategy of their parents unless they are subject to mutation. At the beginning of each generation all players are randomly assigned to unique cliques of the same size. The game is played for many generations to subject the population to selective pressure. We say that cooperation is

established if the average winning strategy (k) for all individuals at the end of the game is less than or equal zero. We find that under our framework, cooperation evolves and is sustained even in larger populations after the game is played for many generations (see Figure Figure 8.2). The effect of the initial clique size, however, is less discernible in larger populations (see the simulation result for n=200 in Figure 8.2).

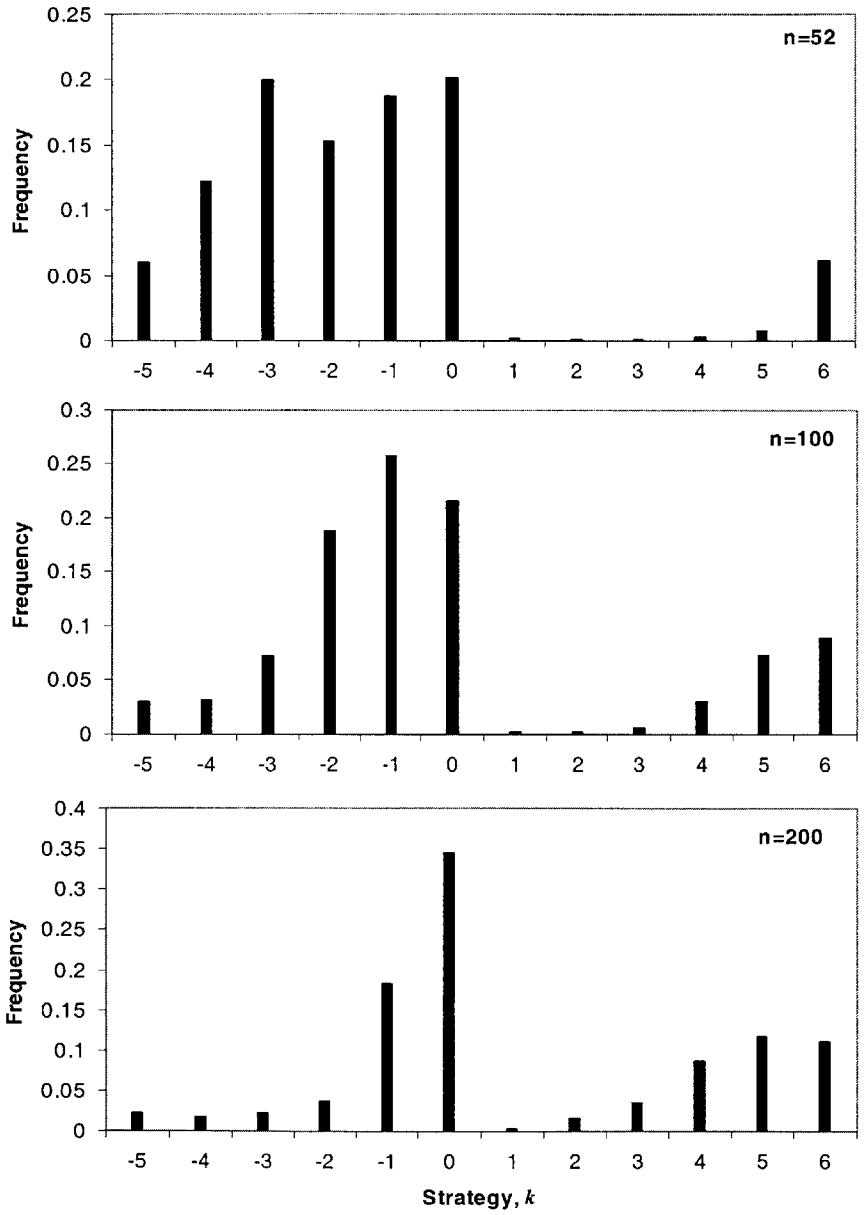


Figure 8.2. Evolution of indirect reciprocity by trust and reputation.

Figure 8.3 shows the result of the simulation for a population of 100 individuals for varying initial clique sizes. With more initial acquaintances cooperation is evolved and sustained in a more secure manner. However, even in an initial population of 'all loners', *i.e.*, when the initial clique size is one, a slight increase in the number of rounds can help a mixture of cooperating strategies to be established. When information is scarce, in the

absence of a dynamically evolving network of acquaintances, cooperators (players with $k \leq 0$) will not stand a chance against defectors. As soon as we allow for channels of information to evolve by letting individuals make new connections as they interact, the likelihood of dissemination of information about players' reputation is increased, thereby increasing the likelihood for discriminators to be able to discriminate rightfully against exploiters, as well as in favor of cooperators. This is why cooperation can evolve under this scenario despite the obvious fact that discriminators are the only ones who can make use of information. Evolution of cooperation is then a consequence of informed discrimination.

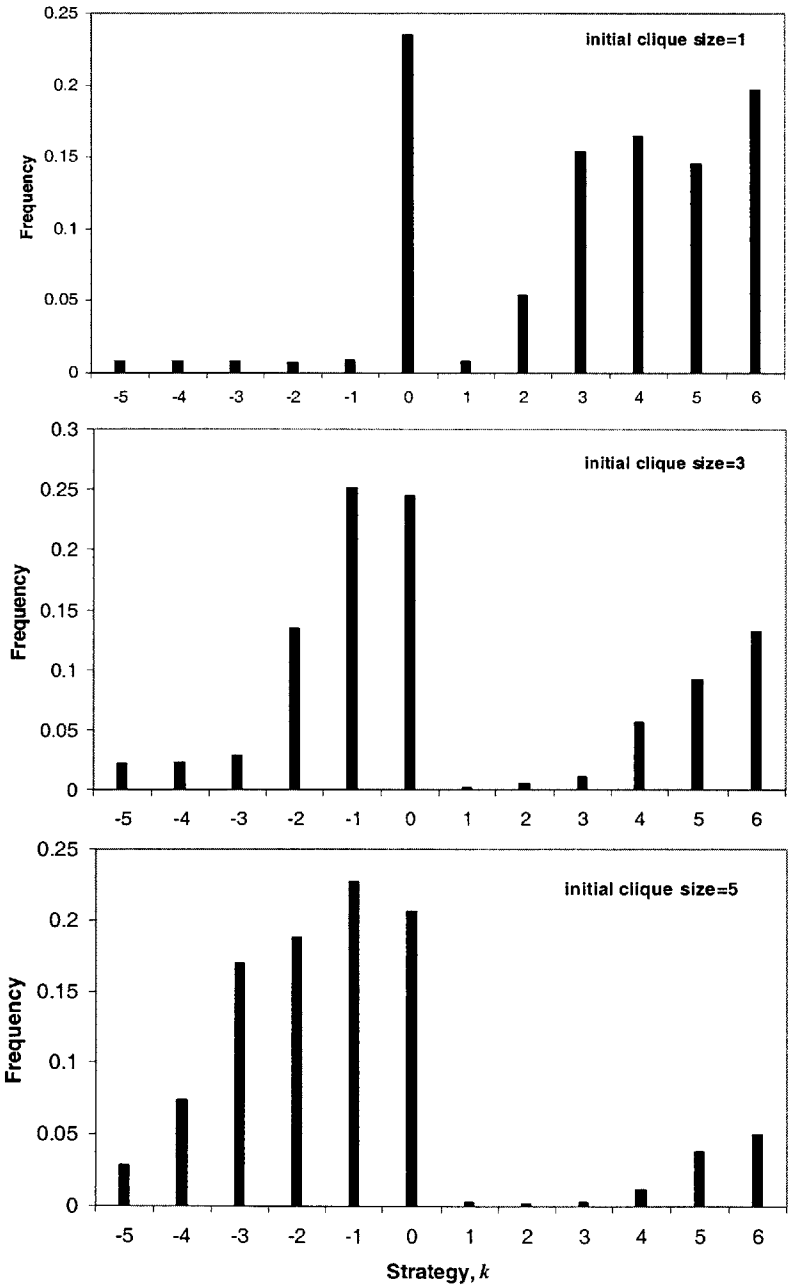


Figure 8.3. Evolution of indirect reciprocity and the initial clique size.

The key hypothesis here is that the ability to compile and process complex social information by way of trust and reputation has played a critical role in the evolution of indirect reciprocity in human societies. Phrasing after Alexander's conjecture that "indirect reciprocity is a consequence of direct reciprocity occurring in the presence of others", we add that indirect reciprocity is also a consequence of inquiring about direct reciprocity.

8.5 Discussion

We have proposed a framework for the evolution of indirect reciprocity by *social information*. Social information is the 'collective memory' and the information content of networks of friends and acquaintances. Such body of information can be retrieved by communicating with friends and acquaintances – as we have modeled:

- Information is selectively retrieved from and propagated through networks of acquaintances due to the existing clustering effect of social ties.
- The topology of the underlying social network is dynamically modified as members changes their interaction patterns.

We have analytically derived the condition under which cooperation can evolve. Our results suggest that for cooperators to be evolutionarily stable, the average benefit received by a trusting and reputably cooperative individual must out-weigh the cost of trusting and helping a disreputable individual. In other words, trust pays off only if it is placed upon the trustworthy.

In this chapter, we have applied our rating propagation from Chapter 3 and 5 to the problem of evolution of indirect reciprocity. We have show how this information retrieved from and propagated through dynamically evolving networks of trust and reputation can have a sustaining effect on the level of cooperation in a community with defectors. We have argued that cooperation is an act of trust and is sustained by reciprocity and propagation of reputation information in a social environment.

CHAPTER 9

Conclusion and Future Work

9.1 What have we learned?

We have provided a critical overview of the state of the art in this field. Many extant studies of trust and reputation studies are made in the context of building reputation or rating systems for online communities. Most of these systems have been constructed without a formal rating model and without much regard to our sociological understanding of these concepts (*e.g.*, Sycara, et al., 1999; Zacharia and Maes, 1999; Yu and Singh, 2000). This is especially true for online reputation or rating systems which claim to encourage trust for their members.

To address these inadequacies, we have first proposed a formal quantitative model for the rating process. Based on this model, we have formulated two personalized rating schemes and have demonstrated their effectiveness at inferring trust experimentally using a simulated dataset and a real world movie-rating dataset. Our experiments show that the popular global rating scheme widely used in commercial electronic communities is inferior to our personalized rating schemes when sufficient ratings among members are available. The level of sufficiency has been discussed.

Secondly, we have proposed a mathematical framework for modeling trust and reputation that is rooted in findings from the social sciences. In particular, our framework makes explicit the importance of social information (*i.e.*, indirect channels of inference) in helping members of a social network choose whom they want to partner with or to avoid.

We argue that a sound reputation or rating system is essential for producing trust in online communities. We have applied our framework to a real world community of movie ratings (MovieLens) and show that our framework can markedly improve the quality of ratings to the users – enabling them to make more trustworthy decisions. In comparison with other models of reputation, we have quantitatively showed that our framework provides significantly better estimations of reputation. "Better" has been discussed with respect to the rating process in Chapter 4 and then to two specific games to be discussed in Chapter 6 and Chapter 8.

Finally, we have extended our trust and reputation framework toward addressing a fundamental problem for social science and biology: evolution of cooperation. We have shown that by providing an indirect inference mechanism for the propagation of trust and reputation, cooperation among selfish agents can be explained for a set of game theoretic

simulations. In particular, our proposal is shown to have provided more cooperative agent communities than existing schemes in the literature are able to.

9.2 Future Work

Although this dissertation has answered how trust and reputation are relevant to cooperation online, it opens up more research opportunities and questions that are unanswered. This section describes a few of these important areas.

9.2.1 Rating Systems

As mentioned toward the end of Chapter 4, a number of unresolved issues regarding rating systems are raised through our work.

Already pointed out is the multiple paths inference problem. Chapter 4 has experimented with several strategies for inference in such setting. We have shown that a good strategy is to use the path that contains the most trusted intermediary. Recent work by Murphy, *et al* (1999), Yedidia, *et al.* (2001) and others have pointed to stochastic techniques for dealing with this multiple paths (or loopy networks) inference problem.

In our formulation, the calculation of an agent's reputation requires the disclosure of detailed personal rating and object descriptions to other agents. This creates a privacy concern, as significant personal information would be contained in this information. Circles of trust could be defined and agents could filter the information they provide to other agents based on their trust of that agent. This does not fully address privacy issues, because trust may be betrayed and circles of trust imply a means to identify agents, allowing agents' activities to be tracked over time or correlated with other data. Addressing this privacy issue is outside the scope of our work here but is an important issue.

Another unresolved issue is the inference mechanism for inferring ratings and reputations from one context to another. We have started investigating the use of ontology to relate different contexts. A hard problem is to determine how to resolve the different ontological views of the world held by different agents. Furthermore, the metric or function to transfer rating or reputation from one context to the next is yet to be worked out. We refer the readers to the paper by Koh and Mui (2001) for our approach toward this problem using the Kullback-Liebler (KL) Divergence measure from information theory.

9.2.2 Trust and Reputation

The study of trust and reputation has been extensive. The written records on their study can be dated back several hundred years, as suggested by the quotes in Chapter 1. Our approach to the study of trust and reputation has been focused on those notions that can implemented in computational programs, and can be applied to enhance the user experience in virtual communities. To reconcile the different notions of trust and

reputation that exist in diverse fields for different scenarios would open up many possible research directions.

In our quantitative comparisons between our reputation propagation framework proposed in Chapter 5 and 6 with other reputation schemes, we have found that our proposal has significant improvement over existing reputation schemes in terms of survival utility to agents in our simulated world. The single stage game played in this simulated world is the Prisoner's Dilemma.

Our results for the Nowak and Sigmund (NS) game in Chapter 8 suggest that our reputation proposal seems to also apply to the NS stage games. Whether the order of strength among the different notions of reputation holds in other types of game can only be speculated on at present. However, our *social information* framework in Chapter 7 and 8 suggests that the more direct and indirect information is available for decision making, the more likely cooperation can evolve. Whether the game is Prisoner's dilemma, Nowak and Sigmund game, or other such common resource games, the amount of reputation information for the interacting agents is likely to contribute to the rise of cooperation.

Our immediate future work is to formulate an information-theoretic framework that maps the amount of "information" available to the likelihood for the evolution of cooperation in a variety of games.

9.2.3 Evolution of Cooperation

We have only scratched the surface with our notion of social information for sustaining cooperation. Many research directions exist for extending our work here. Here are a few:

- Might (simulated or real) societies that judge reputation in different ways simultaneously do better than those with just a single one? In Chapters 4 and 6, we have indirectly shown that by evaluating the reputation information from multiple direct acquaintances, an agent is able to estimate the trustworthiness of an indirect agent better than when only one channel is available.
- Do means of combining reputation from different sources (both types and instances of social networks) lead to different advantages or disadvantages for cooperators?
- Are there significantly different notions of reputation, trust, *etc.*, that apply if one models social interactions by games more complex than the games (Prisoner's Dilemma, and Nowak and Sigmund) that we have used?
- How can we model the actions in society who provide not only information but also enforcement? For example, are there priests who convince their congregation that they will suffer eternally if they are not nice? Or courts that enforce legal contracts? How can these enforcer roles be modeled in some simplified way so as to shed light on the evolution of such institutions and societies?

As computational techniques improve, the study of evolution of cooperation has many promising direction for further exploration.

9.2.4 *Irrationality*

The recent anthropological work by Henrich, *et al.* (2002) comparing how different cultures treat cooperation has challenged much conventional wisdom on rationality. Appendix C outlines how our work is relevant to the modeling of *irrational* behaviors such as cooperation. There is definitely an urgent need for us to take into consideration the cross-cultural variations of cooperative patterns.

9.3 Social Information and Concluding Remarks

Trust and reputation are important sources of information that we gather about each other in our daily lives. This dissertation is an attempt to quantify and formalize some aspects of these two social quantities. Our journey has shown how computational models can be useful in designing rating systems, in explaining cooperation among simulated agents, and in unifying diverse research communities around quantitative frameworks that can be used to benchmark different aspects of these quantities.

As we survey what has been accomplished in this work, we become increasingly aware of the restricted scope in discussing simply these quantities in isolation. The usefulness of trust and reputation perhaps can be understood more profitably by reducing these and other social quantities into some *social information* units.[1] What is important is not about whether a piece of social information is about trust or reputation, but that it contributes to the accumulation of *social information* by the recipients.

With this information-theoretic view, graphs such as that in Figure 6.4 can have their independent axis labeled as "amount of social information". Such an interpretation view the various strategies for gathering trust and reputation for others as no more than different techniques for increasing the *social information* available to the recipients for decision making. How to extend our formulation of social information based on trust and reputation in Chapter 8 to include any generic *social information* will provide a fruitful direction of research.

[1] The spirit of this suggestion is inline with Claude Shannon's work which considers information in the abstract as in binary representation, based upon which much of modern communication theory has been built.

APPENDIX A

Preference based Rating Propagation

The following proof is modified based on a derivation in Ang (2001).

Theorem. With a social network setup in Figure 3.1, the rating propagation function ρ_{ik} when i and k are 2 nodes separated by a third node j is:

$$\rho_{ik} = \begin{cases} \dfrac{(2\rho_i - 1)(2\rho_i\rho_{jk} - \rho_i - \rho_{jk} + \rho_{ij}) + (1-\rho_i)(2\rho_{ij} - 1)}{2\rho_{ij} - 1} & \text{if } \rho_{ij} \neq 0.5 \\ 0.5 & \text{if } \rho_{ij} = 0.5 \end{cases}$$

□

Proof. Given Equation (3.10) is:

$$\rho_{ij}(c) = \rho_i(c)\,\rho_j(c) + (1 - \rho_i(c))\,(1 - \rho_j(c)) \qquad \text{(A.1)}$$

By changing the name of the variables from Equation (3.10), the following are obvious:

$$\rho ik(c) = \rho i(c)\,\rho k(c) + (1 - \rho i(c))\,(1 - \rho k(c)) \qquad (A.2)$$

$$\rho jk(c) = \rho_j(c)\,\rho_k(c) + (1 - \rho_j(c))\,(1 - \rho_k(c)) \qquad \text{(A.3)}$$

For ease of representation, the context variable will be omitted for the rest of the derivation below.

The proof strategy is to derive a closed form for $\rho_{ik}(c)$ based on the above 3 equations so that the only independent quantities are those that are available to i. Since ρ_j is unknown to i, it is a good variable to eliminate. Equation (A.1) yields the following:

$$\begin{aligned}\rho_{ij} &= \rho_i\rho_j + 1 + \rho_i\rho_j - \rho_i - \rho_j \\ &= \rho_j(2\rho_i - 1) - \rho_i + 1 \\ \rho_j &= \frac{\rho_{ij} + \rho_i - 1}{2\rho_i - 1}\end{aligned} \tag{A.4}$$

Similarly, Equation (A.3) yields the following:

$$\begin{aligned}\rho_{jk} &= \rho_j\rho_k + 1 + \rho_j\rho_k - \rho_j - \rho_k \\ &= \rho_j(2\rho_k - 1) - \rho_k + 1 \\ \rho_j &= \frac{\rho_{jk} + \rho_k - 1}{2\rho_k - 1}\end{aligned} \tag{A.5}$$

Setting these 2 ρ_j equal to each other:

$$\begin{aligned}\frac{\rho_{ij} + \rho_i - 1}{2\rho_i - 1} &= \frac{\rho_{jk} + \rho_k - 1}{2\rho_k - 1} \\ (\rho_{ij} + \rho_i - 1)(2\rho_k - 1) &= (\rho_{jk} + \rho_k - 1)(2\rho_i - 1) \\ 2\rho_k(\rho_{ij} + \rho_i - 1) - (\rho_{ij} + \rho_i - 1) &= (\rho_{jk} - 1)(2\rho_i - 1) + \rho_k(2\rho_i - 1) \\ \rho_k(2\rho_{ij} + 2\rho_i - 2 - 2\rho_i + 1) &= (\rho_{jk} - 1)(2\rho_i - 1) + (\rho_{ij} + \rho_i - 1) \\ \rho_k(2\rho_{ij} - 1) &= 2\rho_i\rho_{jk} - \rho_i - \rho_{jk} + \rho_{ij} \\ \rho_k &= \frac{2\rho_i\rho_{jk} - \rho_i - \rho_{jk} + \rho_{ij}}{2\rho_{ij} - 1}\end{aligned} \tag{A.6}$$

With this ρ_k, Equation (A.2) can be expressed in terms of quantities that are available to i :

$$\begin{aligned}\rho_{ik} &= 2\rho_i\rho_k - \rho_i - \rho_k + 1 \\ &= \rho_k(2\rho_i - 1) - \rho_i + 1 \\ &= \frac{(2\rho_i\rho_{jk} - \rho_i - \rho_{jk} + \rho_{ij})(2\rho_i - 1) + (1 - \rho_i)(2\rho_{ij} - 1)}{2\rho_{ij} - 1}\end{aligned} \tag{A.7}$$

By Equation (A.1), the singular point at $\rho_{ij} = 0.5$ implies that $\rho_i = 0.5$. This singular point of the propagation function would yield ρ_{ik}=0.5. This value can be justified by interpreting 0.5 as being the least certain probability. This least uncertainty is warranted since there is no direct and indirect information that i can get about k. Therefore, i is completely uncertain about k.

□

APPENDIX B

Bayesian Rating Propagation

The following derivation first appears in Mui, *et al.* (2001).

This derivation details how Equation (3.28) is arrived at. This derives the posterior estimate of the proportion of approvals in n encounters between individuals a and b is given below:

$$p(\hat{\theta}\,|\,D) = \frac{L(D\,|\,\hat{\theta})p(\hat{\theta})}{\int_{\hat{\theta}} L(D\,|\,\hat{\theta})p(\hat{\theta})d\hat{\theta}}$$

$$= \frac{\hat{\theta}^{p}(1-\hat{\theta})^{n-p}\,\frac{\hat{\theta}^{c_1-1}(1-\hat{\theta})^{c_2-1}}{\int \hat{\theta}'^{c_1-1}(1-\hat{\theta}')^{c_2-1}d\hat{\theta}'}}{\int \hat{\theta}''^{p}(1-\hat{\theta}'')^{n-p}\,\frac{\hat{\theta}''^{c_1-1}(1-\hat{\theta}'')^{c_2-1}}{\int \hat{\theta}'''^{c_1-1}(1-\hat{\theta}''')^{c_2-1}d\hat{\theta}'''}d\hat{\theta}''}$$

$$= \frac{\hat{\theta}^{p+c_1-1}(1-\hat{\theta})^{n+c_1-p-1}}{\int \hat{\theta}''^{p+c_1-1}(1-\hat{\theta}'')^{n+c_1-p-1}d\hat{\theta}''}$$

$$= \frac{\Gamma(c_1+c_2+n)}{\Gamma(c_1+p)\Gamma(n-p+c_2)}\hat{\theta}^{p+c_1-1}(1-\hat{\theta})^{n+c_1-p-1}$$

$$= Beta(c_1+p, c_2+n-p)$$

□

APPENDIX C

Cooperation, Irrationality, and Economics

"The first principle of economics is that every agent is actuated only by self interest."

-- *F. Edgeworth, Mathematical Psychics*

Economists have built the foundation of their discipline on the fundamental assumption that individuals are *rational* and *self-interested.* This is often called the *homo economicus* [1] assumption. Under this assumption, cooperation is puzzling. Individuals have often been found to care about social quantities such as reciprocity, fairness and trust, often at a personal cost to themselves (Henrich, *et al.*, 2002). Through our everyday experiences, we observe that people often take actions to praise and reward those who are cooperative and punish those who are not, even when these actions are costly to the enforcers. This dissertation has suggested the modeling of trust and reputation to explain why rational individuals should be willing to give up what seem to be advantages in the name of cooperation. Cooperation can be understood as rational within our "social information" framework (*c.f.,* Chapter 8). The more the amount of direct and indirect information about agents' history of interaction, the more likely cooperation will evolve for that group of agents.

This appendix provides the background of the argument above by first reviewing the evidences against the rationality claim for cooperation and responses to them in Section C.1. Section C.2 suggests how this dissertation has contributed to this debate and briefly concludes this discussion.

C.1 Irrational Man and Responses

Empirical findings by social scientists have consistently uncovered significant deviations from the predictions of *homo economicus* (Fehr, *et al.*, 2001; Ostrom, 1998; Camerer, 1995; Roth, *et al.*, 1991; Caporael, *et al.*, 1989; Kahneman, *et al.*, 1986). These

[1] The "rational economic man" refers to a number of notions. At its core, it assumes that such an individual choose the "best" among alternatives in a way that "properly" accords with the preferences and beliefs of an individual decision maker or those of a group making a joint decision (Doyle, 1998). "Best" is defined with respect to some maximization operating using well ordered preference relations.

evidence shows that many people are strongly influenced by "moral" preferences, and that concerns for fairness and reciprocity often take precedence over personal gains. Individuals are very willing in engage in cooperative activities even at great risk for personal loss. Anthropologists describe the observed human cooperation as "altruistic" since selfishness-based arguments cannot explain such behaviors (Henrich, *et al.*, 2002).

Socio-biological theories predict that cooperation and altruism should be limited to kin and reciprocating partners (Hamilton, 1963; Trivers, 1971; Axelrod, 1984; Boyd and Richerson, 1989). However, humans cooperate with large groups of unrelated individuals who do not promise reciprocation. Their cooperation is not just co-incidental to their selfish pursuit; anthropological experiments with western subjects have shown that these individuals actually have *social preferences* that support large scale cooperation (Fehr, *et al.*, 2001). Such preferences include: inequality aversion, strong reciprocity, and concerns for fairness.

One of the main goals of the recently completed MacArthur Cross-Cultural Project is to answer whether the canonical selfishness-based models of human decision making holds true across 18 distinct social-economic groups in 4 continents with over 1030 subjects. The results by 12 researchers in economics and anthropology emphatically show that the canonical selfishness-based assumption about human do not explain any of the social groups studied (Henrich, *et al.*, 2002). At the same time, behavioral variability in this well-designed and controlled set of experiments point to a lack of universal pan-human explanation for issues about cooperation and related variables such as reciprocity and trust.

With their fundamental rationality assumption under attack, economists (and rational decision supporters) have three main responses:

- Real world individuals are indeed irrational. However, the aggregate sum of individual decision making that is "rational" – it is the aggregation that economics is concerned with.
- Individuals are boundedly rational. Within the bound of their knowledge and inference mechanisms, they are rational.
- The preference relations of individuals in the real world have not been taking into the model. The failure of the *homo economicus* assumption is therefore a failure of modeling, not one about its premise.

These three arguments are expanded in detail below, along with discussion of weaknesses in their arguments.

C.1.1 Rationality in the Aggregate

Gary Becker (1962, 1974) has formalized individual irrationality into rational aggregation modeling – which has significantly impacted much economic thinking on human behavior.[2] He argues that economic models are about the rational aggregate

[2] For this and his other works on modeling human behavior, Gary Becker won the Nobel Memorial Prize in Economics in 1993.

behavior and not about the irrational individual actions. At the aggregate level, the "average" individual tends to conform to economists' modeling. Such a response is unsatisfactory because it is arguing that limitations of current understanding and analytical tools should be dictating how economic modeling is made.[3] Especially after many claims by micro-economists and other rational decision modelers about how individuals behave in the real world, Becker's claim has the veneer of an unconvincing sleight of hand.

C.1.2 Bounded Rationality

Herbert Simon (1955, 1982) proposes the use of *bounded rationality* to model individual decision making. He admits that living beings are indeed not fully rational in the traditional sense but that within the confines of their knowledge and reasoning faculty, they can be modeled in the same manner. Simon's ideas are very influential, as attested by the ample application of bounded rationality across diverse disciplines (Osborne and Rubinstein, 1998; Gilboa, *et al.*, 1995; Neyman, 1985; *etc.*). Nevertheless, one of the findings in the MacArthur Cross-Cultural Project (among other findings [4]) discussed above is that even when individuals' decision making is confined to simple and well understood set of facts and rules, human behavior can still not be described in ways that are completely selfish and "individually rational." As the 12 researchers have found in this project, factors such as cultures, societal structures and norms have significant influences on how individuals behave beyond considerations of their selfish interests. Note that bounded rationality is not rejected by these experiments; rather, the evidence suggests a reconsideration of its usual behavioral interpretation with self-regarding, exogenous preferences (Henrich, *et al.*, 2002).[5]

C.1.3 Modeling Irrational Preferences

With regard to the third response, many influential economists, including Adam Smith (in his *Theory of Moral Sentiments*, 1759), Kenneth Arrow (1981), Paul Samuelson (1993) and Amartya Sen (1995) have acknowledged that individuals often have "irrational preferences" (*e.g.*, people often care for the well-being of others, sometimes at their own expenses). These economists have noted that such irrationality may have important economic consequences. However, these incidental opinions have little impact in mainstream economic thoughts. As Fehr and Fischbacher (2002) explains, the difficulty for their ideas to gain acceptance is mainly due to a "strong convention in economics of not explaining puzzling observations by changing assumptions on preferences." Economists believe, and rightly so, that by choosing the "right" preferences, everything can be explained.

Economists are unwilling to abandon the *homo economicus* assumption and embrace modeling "irrational preferences" for good reasons. This assumption has

---

[3] The argument here is reminiscent of the tale for finding a missing door key beneath a street lamp and, not in the dark house where the key was lost, since there is light in the street.

[4] Example with similar findings are: Sally (1995), Ostrom (1998), Fehr and Gachter (2000).

[5] In other words, preferences should be endogenously modeled in bounded rationality models and not be taken as exogenous given – domains of anthropologists and behavioral social scientists.

provided good predictions for some important economic problems. In the area of highly competitive markets with standardized goods, this assumption makes excellent predictions[6]. Such markets are nearly always assumed to be in perfect equilibrium. However, much economic activities take place outside such markets. Further, almost no market can be claimed to be in perfect equilibrium at all times.[7]

In fact, the physical spaces that most people interact in cannot be modeled as such. People operate in markets with a small number of traders. These markets are often laden with "informational friction" [8] – which makes fully rational decision difficult (Fehr, *et al.*, 2001). People often work under incompletely specified and incompletely enforceable contracts based on trust and reputation. Individuals often make decisions that are far from well thought-out. Economists term such activities as "irrational behaviors" because they do not conform to conventional thoughts about what is advantageous to the individuals.

C.2 Rational Cooperation

Cooperation has been a puzzle under the traditional theory of rational agency – the underpinning foundation of much of artificial intelligence and economics, and has been called "irrational." This work argues that cooperation can be very rational if the notion of rationality is expanded beyond immediate personal gain. This argument is not novel – of course – as attested by the literature on cooperation already cited in Chapter 7 and this appendix. What is new is our proposition of the concept of *social information* as central to cooperative interaction (*c.f.*, Chapter 8).

Social information enables individuals to refine their interaction behavior based on their preset preferences of cooperative strategies. In Chapter 8, we have shown both analytically and through game theoretic simulations that agents who are able to utilize their social information can acquire more fitness than those who are not.[9] Cooperative behavior can evolve for agents who take advantage of their social information.

Social information is about the social structures such as trust and reputation which are propagated through social networks (Wasserman and Faust, 1994; Granovetter, 1983). The literature on trust and reputation is extensive. To enable an informed research agenda, we have surveyed the literature and have summarized them with a cross-disciplinary perspective in Chapter 2, 3, and 5. We have then provided a sociologically justified, statistically sound computational formulation of trust and reputation in Chapter 5. Quantitative comparisons of our scheme have been performed against existing computation schemes. Our formulation has marked advantages as indicated in results from a set of game-theoretic simulations. In several sets of game theoretic simulations for evaluating the process for the evolution of cooperation, our proposal is shown to have provided more cooperative agent communities. Whereas comparison across different

---

[6] Any economic textbook can provide plenty of examples in this regard. For example, consult Samuelson and Nordhaus (2000).

[7] The recent turmoil in the financial markets in Asia and the stock markets around the world is just but two examples for the "far from equilibrium" claim.

[8] "Informational friction" refers to the lack of transparency about market parameters, individual preferences and payoffs, and uncertainties involving the outcome of any transactions or activities.

[9] Fitness refers to private utility such as wealth (in economic modeling) or number of progeny (in socio-biological modeling).

notions of reputation can only be performed qualitatively before, we provide a quantitative comparison framework for such social quantities.

This appendix has provided some background for how this dissertation is relevant to the irrationality debate. Complete modeling of individual rational and irrational decision making is a very large program and is certainly outside the scope of this work. Our work can be used to study and model why people cooperate, sometimes even at their own expenses.

Our thesis is that cooperation is not irrational. Cooperation can be understood as rational within a "social information" framework. Cooperation can evolve among self-interested individuals if certain social structures are well-established. The elements of social structures studied in this work are trust and reputation.

BIBLIOGRAPHY

P. Abell, D. Reyniers (2000) "Generalized Reciprocity and Reputation in the Theory of Cooperation: a Framework." *Analyse and Kritik, 22*: 3-18.

A. Abdul-Rahman, S. Hailes (2000) "Supporting Trust in Virtual Communities." *33rd Hawaii International Conference on System Sciences.*

G. Akerlof (1970) "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism." *Quarterly Journal of Economics, 84*: 488-500.

R. D. Alexander (1987) *The Biology of Moral System,* New York: Aldine de Gruyter.

J. Andreoni, J. H. Miller (1992) "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence", *The Economic Journal, 103* (418): 570-585.

C. Ang (2001) *A Context-based Personalized Ratings Management System.* Master of Engineering Thesis, Massachusetts Institute of Technology.

R. Axelrod (1984) *The Evolution of Cooperation.* New York: Basic Books.

R. Axelrod, & D. Dion (1988) "The Further Evolution of Cooperation." *Science, 242,* 4884.

R. Axelrod, & W. D. Hamilton (1981) "The Evolution of Cooperation." *Science, 211,* 1390.

A. Baier (1986). "Trust and Antitrust." *Ethics, 96*(2), pp. 231-260.

P. Bajari, A. Hortacsu, (1999) "Winner's Curse, Reserve Prices and Endogenous entry: Empirical Insights from eBay Auctions." *Stanford Institute for Economic Policy Research. SIEPR. Policy paper No. 99-23.*

J. H. Barkow, L. Cosmides, J. Tooby, J. (eds., 1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* Oxford: Oxford University Press

H. Baumgartner, R. Pieters (2000) "The Influence of Marketing Journals: a Citation Analysis of the Discipline and its Sub-Areas." *Center for Economic Research Paper No. 2000-123.* http://citeseer.nj.nec.com/baumgartner00influence.html

L. C. Becker, (1990) *Reciprocity.* Chicago: University of Chicago Press.

G. S. Becker (1962) "Irrational Behavior and Economic Theory." *The Journal of Political Economy, 70* (1): 1-13.

G. S. Becker (1974) "A Theory of Social Interactions." *Journal of Political Economy, 82:* 1063-1093.

K. Binmore (1997) "Rationality and Backward Induction." *Journal of Economic Methodology, 4*: 23-41.

P. Bonacich (1987) "Power and Centrality: A Family of Measures." *American Journal of Sociology, 92*(5): 1170-1182.

R. Boyd, J. Lorberbaum (1987) "No Pure Strategy is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game." *Nature, 327*: 58-59.

R. Boyd, P. J. Richerson (1982) "Cultural Transmission and the Evolution of Cooperative Behavior." *Human Ecology, 10:* 325-351.

R. Boyd, P. J. Richerson (1988) "The Evolution of Reciprocity in Sizeable Groups." *Journal of Theoretical Biology, 132*: 337-356.

C. Camerer (1995) "Individual Decision Making." In J. H. Kagel, A. E. Roth (eds.), *The Handbook of Experimental Economics.* Princeton, NJ: Princeton University Press: 587-703.

L. Caporael, R. M. Dawes, J. M. Orbell, A. J. Van de Kragt (1989) "Selfishness Examined: Cooperation in the Absence of Egoistic Incentives." *Behavioral and Brain Sciences, 12*: 683-697.

C. Castelfranchi, R. Conte, M. Paolucci (1998) "Normative Reputation and the Costs of Compliance." *J. Artificial Societies and Social Simulations, 1*(3).

E. Castillo, J. M. Gutierrez, A. S. Hadi (1997) *Expert Systems and Probabilistic Network Models*, New York: Springer-Verlag.

M. Caullery (1952) *Parasitism and Symbiosis,* London: Sidgwick and Jackson.

M. D. Cohen, R. L. Riolo, R. Axelrod (2001) "The Role of Social Structure in the Maintenance of Cooperative Regimes." *Rationality and Society, 13*: 5-32.

L. Cosmides, J. Tooby (1992) "Cognitive Adaptations for Social Exchange." In Barkow, J. H., Cosmides, L., Tooby, J. (eds.). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, New York: Oxford University Press, 163-228.

P. Dasgupta (2000) "Trust as a Commodity." In Gambetta, D. (ed.) *Trust: Making and Breaking Cooperative* Relations, *electronic edition*, Department of Sociology, University of Oxford.

R. Dawkins (1976) *The Selfish Gene*, Oxford: Oxford University Press.

C. Dellarocas (2000) "Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior." *Proc. 2nd ACM Conference on Electronic Commerce.*

S. Dewan, V. Hsu (2001) "Trust in Electronic Markets: Price Discovery in Generalist Versus Specialty Online Auctions." *http://databases.si.umich.edu/reputations/bib/papers/Dewan&Hsu.doc.*

F. J. Diez (1993) "Parameter adjustment in Bayes networks: The generalized noisy or-gate." In D. Heckerman, A. Mamdani (eds.), *Proc. Ninth Conference on Uncertainty in Artificial Intelligence* (UAI '93), pp. 99--105.

F. J. Diez (1996) "Local conditioning in Bayesian networks." *Artificial Intelligence, 87*, pp. 1-20.

R. Dingledine, M. J. Freedman, D. Molnar (2001) "Free Haven." *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, O'Reilly.

D. Donath (2002) "A semantic approach to visualizing online conversations." *Communications of the ACM, 45*(4).

J. Doyle (1998) "Rational Decision Making." In R. Wilson, F. Kiel (eds.) *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, Massachusetts: MIT Press.

J. Ensminger (2002) "From Nomads in Kenya to Small-Town America: Experimental Economics in the Bush." *Caltech Engineering & Science, 2*: 6-16.

B. Esfandiari, S. Chandrasekharan (2001) "On How Agents Make Friends: Mechanisms for Trust Acquisition." *4th Workshop on Deception, Fraud and Trust in Agent Societies*, Montreal, Canada.

I. Eshel (1972) "On the Neighbourhood Effect and Evolution of Altruistic Traits." *Theor. Population Biology, 3*: 258-277.

E. Fehr, S. Gachter (2000) "Cooperation and Punishment in Public Goods Experiments." *American Economic Review, 90*: 980-994.

E. Fehr, K. Schmidt (2001) "Theories of Fairness and Reciprocity – Evidence and Economic Applications." *Working Paper, No. 75*, Institute for Empirical Research in Economics, University Zurich.

E. Fehr, U. Fischbacher (2002) "Why Social Preferences Matter – the Impact of Non-Selfish Motives on Competition, Cooperation and Incentives." *Working Paper No. 84,* Institute for Empirical Research in Economics, University of Zurich.

R. A. Fisher (1958) The Genetic Theory of Natural Selection. 2nd ed, New York: Dover.

L. C. Freeman (1979) "Centrality in Social Networks: I. Conceptual Clarification." *Social Networks, 1:* 215-239.

J. Friedman (1971) "A Non-cooperative Equilibrium for Supergames." *Review of Economic Studies, 38*: 1-12.

E. Friedman, P. Resnick (1998) "The Social Cost of Cheap Pseudonyms." *Telecommunications Policy Research Conference.*

D. Fudenberg, E. Maskin (1986) "The Folk Theorem in Repeated Games with Discounting and Incomplete Information." *Econometrica, 54*: 533-554.

D. Fudenberg, J. Tirole (1991) *Game Theory*, Cambridge, Massachusetts: MIT Press.

D. Gambetta (1988) Trust: Making and Breaking Cooperative Relations, Oxford: Basil Blackwell.

E. Garfield (1955) "Citation Indexes for Science." *Science, 122*: 108-111.

I. Gilboa, D. Schmeidler (1995). "Case-based Decision Theory." *Quarterly Journal of Economics, 110*: 605-639.

A. Glass, B. Grosz (2000) "Socially Conscious Decision-Making." *Proc. Autonomous Agents'2000.*

A. W. Gouldner (1960) "The Norm of Reciprocity: A Preliminary Statement." *American Sociological Review, 25*: 161-78.

D. Goldberg, D. Nichols, B. M. Oki, D. Terry (1992). "Using Collaborative Filtering to Weavean Information Tapestry." *Communications of the ACM, Dec.*

M. Granovetter (1985) "Economic Action and Social Structure: the Problem of Embeddedness." *American Journal of Sociology, 91*: 481-510.

J. W. Grossman (1997) "Paul Erdös: The Master of Collaboration." In R. L. Graham, J. Nesetril (eds.) *The Mathematics of Paul Erdös*, Springer-Verlag.

W. Guth, H. Kliemt (1998) "The Indirect Evolutionary Approach: Bridging the Gap between Rationality and Adaptation." *Rationality and Society, 10*(3): 377 – 399.

A. Halberstadt, L. Mui (2001) "Group and Reputation Modeling in Multi-Agent Systems." *Proc. Goddard/JPL Workshop on Radical Agents Concepts, NASA Goddard Space Flight Center.*

J. B. S. Haldane (1932) *The Causes of Evolution.* London: Longmans Geen & Co.

W. D. Hamilton, (1963). "The Evolution of Altruistic Behavior." *American Naturalist. 97*: 354-356.

W. D. Hamilton, (1964). "The Ggenetical Evolution of Social Behavior." *J. Theor. Biol. 7*: 1-16.

G. Hardin, (1968) "The Tragedy of the Commons." *Science 162*(1): 243-48.

R. Hardin (1997) "Economic Theories of the State." In D. C. Mueller. (ed.) *Perspectives on Public Choice: A Handbook*, Cambridge: Cambridge University Press: 21-34.

J. Harsanyi (1967) "Games with Incomplete Information Played by Bayesian Players." *Management Review, 14*: 159-182, 320-334, 486-502.

D. Heckerman (1996). "A Tutorial on Learning with Bayesian Networks." *Technical Report MSR-TR-95-06*, Microsoft Research.

J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath (2002) "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *Economics and Social Behavior, 91* (2): 73-78.

E. J. Horvitz (1986) "Reasoning about Beliefs and Actions under Computational Resource Constraints." *Proceedings of the Third AAAI Workshop on Uncertainty in Artificial Intelligence.*

D. E. Houser and J. Wooders (2001) "Reputation in Internet Auctions: Theory and Evidence from eBay." working paper: *http://w3.arizona.edu/~econ/working_papers/*Internet_Auctions.pdf.

D. Hume (1737) *A Treatise of Human Nature*. L. A. Selby-Bigge, P. H. Nidditch (eds). Oxford: Clarendon Press, 1975.

D. Kahneman, J. Knetsch, R. Thaler (1986) "Fairness as a Constraint on Profit-Seeking: Entitlements in the Market." *American Economic Review 76*, 4, p. 728-741.

M. Kandori (1992) "Social Norms and Community Enforcement." *The Review of Economic Studies, 59* (1): 63-80.

M. Kandori (2002) "Introduction to Repeated Games with Private Monitoring." *Journal of Economic Theory, 102*: 1-15.

L. Katz (1953) "A New Status Index Derived from Sociometric Analysis." *Psychometrika, 18*: 39-43.

R. L. Keeney, H. Raiffa (1976) Decisions With Multiple Objectives: Preferences and Value Tradeoffs, New York: John Wiley and Sons.

L. Keller, H. K. Reeve (1998) "Familiarity Breeds Cooperation." *Nature, 394:* 121-122.

R. Khare, A. Rifkin (1997) "Weaving a Web of Trust." *World Wide Web Journal, 2*(3): 77-112.

W. Koh, L. Mui (2001) "An Information-Theoretic Approach for Ontology-based Interest Matching." International Joint Conference on Artificial Intelligence (IJCAI) Ontology Learning Workshop, Seattle, WA.

P. Kollock (1994) "The Emergence of Exchange Structures: An Experimental Study of Uncertainty, Commitment, and Trust." *American Journal of Sociology, 100*(2): 313-345.

P. Kollock, M. Smith (1996) "Managing the Virtual Commons: Cooperation and Conflict in Computer Communities." S. Herring (ed.), *Computer-Mediated Communication*, Amsterdam: John Benjamins, 1996.

D. Krackhardt, M. Lundberg, L. O'Rourke (1993) "KrackPlot: A Picture's Worth a Thousand Words." *Connections, 16*: 37-47.

W. H. Krackle (1978) *Force and Persuation: Leadership in an Amazonian Society,* Chicago, IL: University of Chicago Press.

D. Kreps (1990) *A Course in Microeconomics*. Princeton, NJ: Princeton University Press.

D. M. Kreps, R. Wilson (1982) "Reputation and Imperfect Information." *Journal of Economic Theory, 27*: 253-279.

Y. Lashkari, M. Metral, P. Maes (1994) "Collaborative Interface Agents." In *Proceedings of the 12th National Conference on Artificial Intelligence.*

M. P. Lombardo (1985) "Mutual Restraint in Tree Swallows – a Test of the Tit for Tat Model of Reciprocity." *Science, 227* (4692): 1363-1365.

A. Lotem, M. A. Fishman, L. Stone (1999) "Evolution of Cooperation between Individuals." *Nature, 400*: 226-227.

D. Lucking-Reiley, D. Bryan, N. Prasa, D. Reeves (1999) "Pennies from eBay: The Determinants of Price in Online Auctions." http://eller.arizona.edu/ ~reiley/ papers/PenniesFromEBay.pdf

K. A. McCabe, S. J. Rassenti, V. L. Smith (1996) "Game Theory and Reciprocity in Some Extensive Form Experimental Games." *Proceedings of the National Academy of Sciences*: 9313421-9313428

J. Makino, Y. Fujigaki, and Y. Imai (1997) "Productivity of Research Groups – Relation between Citation Analysis and Reputation within Research Community." *Japan Journal of Science, Technology and Society, 7*: 85-100.

P. V. Marsden, N. Lin (eds.) *Social Structure and Network Analysis,* Newbury Park, CA: Sage.

S. Marsh (1994) *Formalising Trust as a Computational Concept.* Ph.D. Thesis, University of Stirling.

J. Maynard Smith, G. R. Price (1973) "The Logic of Animal Conflict." *Nature, 246*: 15-18.

J. Maynard Smith (1982) *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.

P. R. Milgrom, D. C., North, B. R. Weingast (1990) "The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs." *Economics and Politics*, *2*(1): 1-23.

P. R. Milgrom, J. Roberts (1982) "Predation, Reputation and Entry Deterrence." *Journal of Economic Theory, 27*: 280-312.

M. Milinski (1987) "Tit-for-tat in Sticklebacks and the Evolution of Cooperation." *Nature, 325*: 433-435.

L. Mui, C. Ang, M. Mohtashemi, P. Szolovits, A. Halberstadt (2001) "Collaborative Sanctioning : Enabling Principled Reliability and Reputation Ratings for Distributed Systems." *Laboratory for Computer Science Technical Memorandum No. 617*, Massachusetts Institute of Technology.

L. Mui, M. Mohtashemi, C. Ang, P. Szolovits, A. Halberstadt, A. (2001) "Ratings in Distributed Systems: A Bayesian Approach." *11th Workshop on Information Technologies and Systems (WITS)*, New Orleans.

L. Mui, M. Mohtashemi, A. Halberstadt (2002) "A Computational Model for Trust and Reputation." *35th Hawaii International Conference on System Sciences.*

L. Mui, M. Mohtashemi (2002). "Rational Decision Making Using Social Information." *MIT LCS Technical Report.*

K. P. Murphy, Y. Weiss, M. I. Jordan (1999). "Loopy belief propagation for approximate inference: an empirical study." *Proceedings of Uncertainty in AI.*

M. G. Murray, R. J. Gerrard (1984) "Conflict in the Neighbourhood: Models where Close Relatives are in Direct Competition." *Journal of Theoretical Biology, 111*: 237-246.

A. Neyman (1985) "Bounded Complexity Justifies Cooperation in the Finitely Repeated Prisoner's Dilemma." *Economics Letters, 19:* 227-229.

M. A. Nowak, and K. Sigmund (1998) "Evolution of Indirect Reciprocity by Image Scoring." *Nature, 393*: 573-577.

M. A. Nowak, & K. Sigmund, (1998). "The Dynamics of Indirect Reciprocity." *J. Theor. Biol. 194*,561-574.

M. A. Nowak, and K. Sigmund (2000) "Cooperation versus Competition." *Financial Analyst Journal, July/August*: 13-22.

M. Okuno-Fujiwara, A. Postlewaite (1995) "Social Norms and Random Matching Games." *Games and Economic Behavior, 9*: 79-109.

M. J. Osborne, A. Rubinstein (1998) "Games with Procedurally Rational Players." *American Economic Review, 88*: 834-847.

E. Ostrom (1998) "A Behavioral Approach to the Rational-Choice Theory of Collective Action." *American Political Science Review, 92*(1): 1-22.

G. B. Pollock, L. A. Dugatkin (1992) "Reciprocity and the Evolution of Reputation." *Journal of Theoretical Biology, 159*: 25-37.

D. C. Queller (1994) "Genetic Relatedness in Viscous Populations." *Evolutionary Ecology, 8*: 70-73.

A. Rapoport (1997) "Order of Play in Strategically Equivalent Games in Extensive Form." *International* Journal *of Game Theory, 26*(1): 113-136.

W. Raub, J. Weesie (1990) "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology, 96*(3): 626-654.

C. Renfrew, P. Bahn (1996) *Anchaeology: theories, Methods, and Practice.* London: Thames & Hudson.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl (1994) "GroupLens: An Open Architecture for Collaborative Filtering of Netnews." In Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work (CSCW).

P. Resnick , K. Kuwabara, R. Zeckhauser , E. Friedman (2000a) "Reputation Systems." *Communications of the ACM, 43*(12): 45-48.

P. Resnick, R. Zeckhauser (2000b) "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputatoin System." *NBER Workshop on Empirical Studies of Electronic Commerce Paper.*

H. Rheingold (2001) The Virtual Community : Homesteading on the Electronic Frontier, Revised Edition. Cambridge, MA: MIT Press.

R. L. Riolo, M. D. Cohen, R. Axelrod (2001) "Evolution of Cooperation without Reciprocity." *Nature, 414:* 441-443.

G. Roberts, T. N. Sherratt (1998) "Development of Cooperative Relationships through Increasing Investment." *Nature, 394*: 175-179.

S. Ross (1995) *Stochastic Processes.* John Wiley & Sons.

A. E. Roth, V. Prasnikar, M. Okuno-Fujiwara, A. Zamir (1991) "Bargining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review, 81* (5): 1068-1095.

J. Rouchier, M. O'Connor, F. Bousquet (2001) "The Creation of a Reputation in an Artificial Society Organized by a Gift System." *Journal of Artificial Societies and Social Simulations, 4*(2).

J. Sabater, C. Sierra (2001) "REGRET: A reputation Model for Gregarious Societies." *4th Workshop on Deception, Fraud and Trust in Agent Societies.*

D. Sally (1995) "Conversation and Cooperation in Social Dilemmas: a Meta-Analysis of Experiments from 1958 to 1992." *Rationality and Society, 7* (1): 58-92.

P. A. Samuelson, W. D. Nordhaus (2000) *Economics*, McGraw-Hill.

P. A. Samuelson (1993) "Altruism as a Problem Involving Group versus Individual Selection in Economics and Biology." *American Economic Review, 83:* 143-148.

M. Schillo, P. Funk, M. Rovatsos (2000) "Using Trust for Detecting Deceitful Agents in Artificial Societies." *Applied Artificial Intelligence, Special Issue on Trust, Deception and Fraud in Agent Societies.*

R. Selten (1978) "The Chain Store Paradox." *Theory and Decision, 9*: 127-159.

A. Sen (1995) "Moral Codes and Economic Success." In C. S. Britten, A. Hamlin (eds.) *Market Capitalism and Moral Values*, Edward Eldar, Aldershot.

H. Simon (1969) *The Science of the Artificial,* Cambridge, MA: MIT Press.

H. Simon (1990) "A Mechanism for Social Selection and Successful Altruism." *Science 250*: 1665-1668.

H. Simon (1991) "Organizations and Markets." *J. Econ. Perspectives, 5*: 859-871.

A. Smith (1759, reprint 1982) *The Theory of Moral Sentiments*, Indianapolis: Liberty Fund.

M. Smith and P. Kollock (eds., 1999) *Communities in Cyberspace.* London: Routledge Press.

G. Strang (1988) *Linear Algebra and its Applications.* San Diego: Harcourt Brace and Jovanovich Publishers.

K. Sycara, J. Lu, M. Klusch, and S. Widoff (1999) "Matchmaking among Heterogeneous Agents on the Internet." In Proceedings of the 1999 AAAI Spring Symposium on Intelligent Agents in Cyberspace.

S. Tadelis (1999) "What's in a Name? Reputation as a Tradeable Asset." *American Economic Review, 89*(3): 548-563.

S. Tadelis (2000) "Firm Reputation with Hidden Information." *Stanford Economics Working Paper.*

P. D. Taylor (1992) "Inclusive Fitness in a Homogeneous Environment." *Proc. R. Soc. London. B, 249:* 299-302.

J. Tirole (1996) "A Theory of Collective Reputation (with Applications to the Persistence of Corruption and to Firm Quality." *The Review of Economic Studies, 63*(1): 1-22.

R. L. Trivers, (1971) "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology, 46*: 35-57.

U. S. Department of Justice (2001) *Press Release.* http://www.usdoj.gov/criminal/cybercrime/ebayplea. htm

M. Waldman, A. D. Rubin, L. F. Cranor (2000) "Publius: A Robust, Tamper-Evident, Censorship-Resistent Web Publishing System." *Proc. 9th USENIX Security Symposium.*

S. Wasserman, K. Faust (1994) *Social Network Analysis: Methods and Applications.* Cambridge University Press.

C. Wedekind, M. Milinski (2000) "Cooperation through Image Scoring in Humans." *Science, 288*: 850-852.

M. P. Wellman (1993) "A Market-oriented Programming Environment and its Application to Distributed Multicommodity Flow Problems." *Journal of Artificial Intelligence Research, 1*: 1-23.

B. Wellman (2001) "Computer Networks As Social Networks." *Science 293*(14), pp. 2031-2034.

S. A. West, M. G., Murray, C. A. Machado, A. S. Griffin, E. A. Herre (2001) "Testing Hamilton's Rule with Competition between Relatives." *Nature, 409*: 510-513.

G. C. Williams (1971). *Group Selection.* Aldine-Atherton, Chicago.

D. S. Wilson (1980) *The Natural Selection of Populations and Communities.* Menlo Park, CA: Benjamin Cummings.

D. S. Wilson, G. B. Pollock, L. A. Dugatkin (1992) "Can Altruism Evolve in Purely Viscous Populations." *Evolutionary Ecology, 6*: 331-341.

D. S. Wilson, E. Sober (1994). Reintroducing group selection to the human behavioral sciences. *Behavior and Brain Science, 17*: 585-654.

J. Yedidia, W. T. Freeman, Y. Weiss (2001) "Generalized Belief Propagation." In T. K. Leen, T. G. Dietterich, V. Tresp. *Advances in Neural Information Processing,* pp. 689-695.

B. Yu, M. P. Singh (2001). "Towards a Probabililstic Model of Distributed Reputation Management." *4th Workshop on Deception, Fraud and Trust in Agent Societies,* Montreal, Canada.

G. Zacharia, P. Maes (1999). "Collaborative Reputation Mechanisms in Electronic Marketplaces." *Proc. 32nd Hawaii International Conf on System Sciences.*

P. R. Zimmerman (1995) *The Official PGP User's Guide*, Cambridge, Massachusetts: MIT Press.