Trust and Nuanced Profile Similarity in Online Social Networks

Jennifer Golbeck[*] golbeck@cs.umd.edu
University of Maryland, College Park
8400 Baltimore Avenue, Suite 200, College Park, Maryland 20740

Abstract

Online communities, where users maintain lists of friends and express their preferences for items like movies, music, or books, are very popular. The web-based nature of this information makes it ideal for use in a variety of intelligent systems that can take advantage of the users' social and personal data . For those systems to be effective, however, it is important to understand the relationship between social and personal preferences. In this work we investigate features of profile similarity and how those relate to the way users determine trust. Through a controlled study, we isolate several profile features beyond overall similarity that affect how much subjects trust a hypothetical users. We then use data from FilmTrust, a real social network where users rate movies, and show that the profile features discovered in the experiment allow us to more accurately predict trust than when using only overall similarity. In this paper, we present these experimental results and discuss the potential implications for social networking and intelligent systems.

1. Introduction

Social notions of trust have become a topic of interest among the computer science community. Trust provides information about with whom we should share information, from whom we should accept information, and what consideration to give to information from people when aggregating or filtering data. At the same time, social networking has become a major movement on the web, with hundreds of websites comprising hundreds of millions of users (Golbeck, 2005). Many social networks allow users to express information about their relationships, including how much they trust people. This provides a minimally invasive way to collect information about users' social contexts. By using information that users are already expressing, there is little overhead to gathering trust data which can then be analyzed and used to create socially intelligent systems.

To effectively use trust in systems it is important to understand what users mean when they say that they trust someone and how much they trust them. The sociological literature has studied reasons for trust extensively. However, when using web-based social networks, most of the information that sociology considers important is not available (e.g. we do not know the history between people, the user's own background and how likely they are to trust in general, the familial/business/friend relationship between users, etc.). Thus, we must understand trust from only the available information.

In the case of trust in web-based social networks, the information we have available is the social network, the trust values in it, and the profiles of users. Those profiles include personal information and frequently include the users' opinions and ratings of items. This

[*]. Corresponding author.

information can be used to explain the magnitude of trust between people or to compute recommendations about how much one user should trust another. There are many algorithms that use only the social network and trust values to compute how much one user should trust another. These have been shown to give relatively accurate results, but are only effective when users are connected in the social network. Collaborative filtering (CF) algorithms, on the other hand, use only profile information when computing recommendations for users. CF algorithms generally compute the overall similarity or correlation between users, and use that as a weight when making recommendations.These algorithms are relatively good at making recommendations for users, and could be applied to computing trust recommendations, but they are only effective when users have a common set of rated items. When the ratings are sparse, it is difficult to compute similarity measures between users.

How profile similarity relates to trust is a relatively unexplored space. Trust-based recommender systems like (Ziegler, 2005), (Massa & Bhattacharjee, 2004), (Golbeck, 2006) assume trust captures similarity between users. The relationship between trust and overall similarity was shown in (Ziegler & Golbeck, 2006). However, in previous work, we saw that trust-based recommendations *outperformed* collaborative filtering algorithms in certain cases (Golbeck, 2005). Because collaborative filtering algorithms use overall similarity of user profiles to make recommendations, those results suggested that when users assign trust, they are capturing more than just overall similarity.

In this work, we explore the relationship between trust and profile similarity. We will show through surveys and analysis of data in existing systems that when users express trust, they are capturing many facets of similarity with other users. We begin by presenting a definition of trust and related work on similarity, trust, and trust computation. We then present a study where users are given generated profiles for hypothetical users, and from the results we extract several features of profile similarity that correlate with trust. Those results are then brought to data taken from our existing FilmTrust system, and we show that using that set of features to predict trust is better correlated with known trust values, and is more accurate than using overall similarity alone. Finally, we discuss the implications of this work for creating intelligent systems that respect users' social preferences.

2. Background and Related Work

As a social concept, trust has many facets and influences that are beyond the scope of what can be captured in social networks or online systems. To compute with trust, it is important to know precisely what is being captured and what its features are. In this section, we present a definition of trust and related work on how people assign trust in general, and in online systems.

2.1 A Definition of Trust

Social trust depends on a host of factors which cannot be easily modeled in a computational system. Past experience with a person and with their friends, opinions of the actions a person has taken, psychological factors impacted by a lifetime of history and events (most completely unrelated to the person we are deciding to trust or not trust), rumor, influence by others' opinions, and motives to gain something extra by extending trust are just a few of these factors. For trust to be used as a rating between people in social networks, the definition must be focused and simplified.

(Marsh, 1994) addressed the issue of formalizing trust as a computational concept in his PhD dissertation at the University of Stirling. His model is complex and based on social and psychological factors. Although this work is often cited, the model is highly theoretical and difficult to implement. It is particularly inappropriate for use in social networks because his focus was on interacting agents that could maintain information about history and observed behaviors. In all web-based social networks that implement trust, users assign a single rating without explicit context or history to their neighbors and thus much of the information necessary for a system like Marsh's is missing. To capture the nuanced information required by an implementation of Marsh's system, users would need to log tremendous amounts of information, and even then the complexity of data required by Marsh's model would be unsatisfied. One of the stated goals of our work is to take advantage of social networks as they exist on the web with users adding simple expressions of trust as they determine it (rather than computing it for them). This differs from the agent system in Marsh's work, and that divergence in goals prevents an application of his model in this context.

(Castelfranchi & Falcone, 1998), (Castelfranchi & Falcone, 2002) present an excellent analysis of trust in multi-agent systems. The break trust down into its components, including the beliefs that must be held to develop trust, how that relates to previous experience, as well as giving descriptions of when trust is rational and irrational. Their work draws largely on the psychological literature, and includes many psychological factors in developing a model for trust in multi agent systems. This approach, however, does not carry over into our work with web-based social networks. Our premise is to utilize only the explicitly stated information that users are already providing within these websites, and that does not include background information, nor does it track historical interactions. In this work, we want to show that useful results can be produced without this extensive background. Because the data that we have chosen as our focus is limited to single values, the more advanced and complex model of Castelfranchi and Falcone does not translate into our case.

(Deutsch, 1962) contains a frequently referenced definition of trust. He states that trusting behavior occurs when a person (say Alice) encounters a situation where she perceives an ambiguous path. The result of following the path can be good or bad, and the occurrence of the good or bad result is contingent on the action of another person (say Bob). Furthermore, the negative impact of the bad result is greater than the positive impact of the good result. This further motivates Alice to make the correct choice. If Alice chooses to go down the path, she has made a trusting choice. She trusts that Bob will take the steps necessary to ensure the good outcome. The requirement that the bad outcome must have greater negative implications than the good outcome has positive implications has been countered in other work (Golebmiewski & McConike, 1975), which does not always require disparity.

(Sztompka, 1999) presents and justifies a simple, general definition of trust similar to that of Deutsch: “Trust is a bet about the future contingent actions of others.” There are two main components of this definition: belief and commitment. First, a person believes that the trusted person will act in a certain way. The belief alone, however, is not enough to say there is trust. Trust occurs when that belief is used as the foundation for making a commitment to a particular action. These two components are also present in the core of

Deutsch's definition: we commit to take the ambiguous path if we believe that the trusted person will take the action that will produce the good outcome.

We adopt this as the definition of trust for our work: trust in a person is a commitment to an action based on a belief that the future actions of that person will lead to a good outcome. The action and commitment does not have to be significant. We could say Alice trusts Bob regarding email if she chooses to read a message (commits to an action) that Bob sends her (based on her belief that Bob will not waste her time). When subjects in our experiments and users of our systems are asked to assign trust, they are given this definition.

2.2 Trust and Similarity

The sociology and social psychology literature extensively addresses factors in trust, but the coverage does directly address how trust relates to similarity (overall or in more specific ways). There is, however, strong results linking similarity and friendship, or interpersonal attraction. A positive relationship between attitude similarity and friendship has been shown in (Burgess & Wallin, 1943), (Newcomb, 1961), (Byrne, 1961), and (Byrne, 1971).

In the recommender systems literature, there have been several studies that touch on similarity and trust. (Sinha & Swearingen, 2001) and (Swearingen & Sinha, 2001) have shown that users tend to trust systems more if they recommend items that they previously liked. That is, a system that demonstrates similarity with the user's preference is more trustworthy. Users also tended to prefer recommendations from friends, and the authors suggest that this is because people have more trust for friends.

A connection between user similarity and trust was established in (Ziegler & Golbeck, 2006). Through an analysis of two systems, this work showed that there was a strong and significant correlation between trust and similarity; the more similar two people were, the greater the trust between them.

The correlation of trust and similarity is not surprising. However, does trust *only* capture overall similarity between people? Clearly it reflects other social factors. Is is possible that trust also captures more nuanced facets of correlation between users in online systems? In related work on the FilmTrust system we found results that suggest users' trust ratings do capture more than just overall similarity. In section 3 we present those results and describe how they have motivated this work, which looks further into how similarity affects trust.

2.3 Computing Trust from Trust

In a system that has a trust component, users will have made some direct statements about people they trust. These statements form a social network. In this paper we are examining how trust corresponds to different features of profile similarity, but if we have direct trust information from the users, that can be a more effective means of predicting how much one person will trust another.

In fact, there is a large body of work on algorithms for inferring trust in social networks. While designed for peer-to-peer systems rather than social networks, one of the most widely cited trust algorithms is EigenTrust (Kamvar, Schlosser, & Garcia-Molina, 2004). It considers trust as a function of corrupt vs. valid files that that a peer provides. A peer maintains information about the trustworthiness of peers with which it has interacted based on the proportion of good files it has received from that peer. For one peer to determine the trustworthiness of another with which it has not interacted, it needs to gather information from the network and infer the trustworthiness. The EigenTrust algorithm calculates trust with a variation on the PageRank algorithm (Page, Brin, Motwani, & Winograd, 1998), used by Google for rating the relevance of web pages to a search.

Advogato is a website, at http://advogato.org, that serves as a community discussion board and resource for free software developers. It also is the testbed for Raph Levin's trust metrics research (Levin & Aiken, 1998). Each user on the site has a single trust rating calculated from the perspective of designated seeds (authoritative nodes). Trust calculations are made using a network flow model. His metric composes certifications between members to determine the trust level of a person, and thus their membership within a group. Users can be certified at three levels: apprentice, journeyer, and master.

(Ziegler & Lausen, 2004) propose a trust algorithm called Appleseed. Like Advogato, it is a group trust metric. However, instead of using maximum flow, the basic intuition is motivated by spreading activation strategies. Like EigenTrust, Appleseed is based on finding the principal eigenvector. It is a local trust metric, and given a network and a source it returns a ranking of all the nodes in the network.

One problem that arises in algorithms that are based on finding the principal eigenvector, like (Kamvar et al., 2004), (Ziegler & Lausen, 2004), and (Richardson, Agrawal, & Domingos, 2003), is that trust must first be normalized to work within the matrix. This means that the normalized trust value from a person who has made many trust ratings will be lower than if only one or two people had been rated. However, socially, trust is not a finite resource; it is possible to have very high trust for a large number of people, and that trust is not any weaker than the trust held by a person who only trusts one or two others.

(Golbeck, 2005) introduces the TidalTrust algorithm for computing personalized trust values in social networks. Unlike the eigenvalue-baed approaches, it outputs a trust recommendation in the same scale that users assign trust values. The details of this algorithm are discussed further in section 3.2.1.

These algorithms are generally quite effective at predicting trust and for improving recommendations in certain cases. However, there are drawbacks that make them inapplicable in many cases. Specifically, they can only be used when there are paths in the social network connecting individuals *and* when the trust values on those paths are accessible. In some cases, paths between two people will not exist in the network. In our own system, FilmTrust, over 58% of the users have *no* friends in the system, and over 60% are connected only into small clusters, detatched from the main component (see Figure 1). Thus, for the majority of users, network-based trust algorithms will be ineffective. Also, while the FilmTrust system uses the trust values of any user to compute recommendations, privacy is always a concern. It is reasonable to assume that many users or networks will not allow a system to access trust values, directly or indirectly. In that case, the only trust information available is the set of values assigned directly by the user

3. FilmTrust Website

In earlier work we developed the FilmTrust website, a trust-based social networking website where users can rate and review movies. We used trust in that system to produce predictive

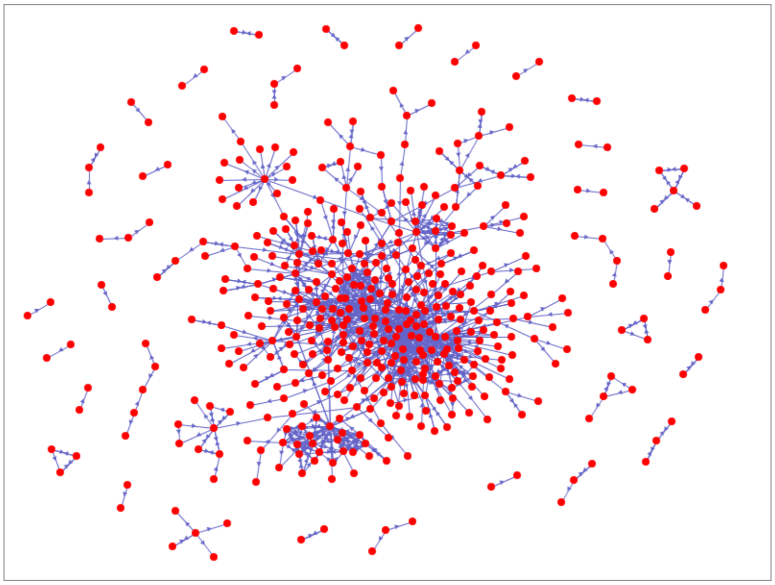
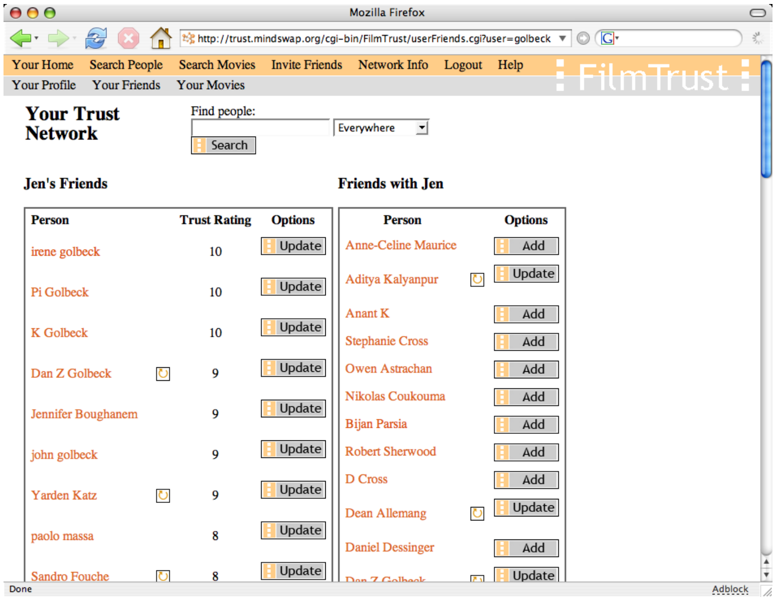


Figure 1: The FilmTrust social network. Users with no friends are not shown, although they constitute 58% of the population. Other users are not connected into the main component and are shown in the small clusters scattered around the edges.



Your Trust Network

Jen's Friends

Person		Trust Rating	Options
irene golbeck		10	Update
Pi Golbeck		10	Update
K Golbeck		10	Update
Dan Z Golbeck	↻	9	Update
Jennifer Boughanem		9	Update
john golbeck		9	Update
Yarden Katz	↻	9	Update
paolo massa		8	Update
Sandro Fouche	↻	8	Update

Friends with Jen

Person		Options
Anne-Celine Maurice		Add
Aditya Kalyanpur	↻	Update
Anant K		Add
Stephanie Cross		Add
Owen Astrachan		Add
Nikolas Coukouma		Add
Bijan Parsia		Add
Robert Sherwood		Add
D Cross		Add
Dean Allemang	↻	Update
Daniel Dessinger		Add
Dan Z Golbeck	↻	Update

Figure 2: A sample Friends page from FilmTrust. Friends of the user are on the left, with the user's trust ratings. People who have listed the user as a friend are on the right.

movie ratings, and the interesting results that we obtained in our analysis motivated this work that further examines trust and user similarity. We will also use data from FilmTrust in the analysis in section 6.

In this section, we introduce the FilmTrust website and present an overview of the relevant results. More thorough discussions of this system and the results can be found in (Golbeck, 2005), (Golbeck, 2006), and (Golbeck & Hendler, 2006).

3.1 The FilmTrust Website

The FilmTrust website, at http://trust.mindswap.org/FilmTrust, is a social network with trust ratings. Users maintain the standard social network information, and also rate movies and write reviews.

In the social networking part of the website, users make lists of their friends. They also are required to assign a trust rating to each friend that indicates how much they trust that person's opinions about movies. Trust is asymmetric - people do not necessarily trust each other the same amount - and we believe friendship relationships are also asymmetric. Thus, relationships are not required to be mutual in our system. Users keep a list of friends, but the friends to not approve the relationship. Thus, on the user's "Friends" page (see figure 2), there are two lists: the people the user has listed as friends (with trust ratings), and the people who have listed the user as a friend. Because of the sensitive nature of trust

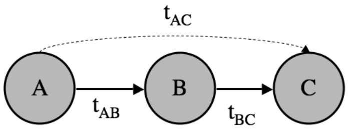


Figure 3: An illustration of direct trust values between nodes A and B (t_{AB}), and between nodes B and C (t_{BC}). Using a trust inference algorithm, it is possible to compute a value to recommend how much A may trust C (t_{AC}).

ratings, the user is the only person who can see the trust values they assign. These values are hidden from everyone else.

In the movies section of their space, users can rate and review films. Movie ratings are on a scale where a half star is the worst rating, and 4 stars is the best rating. Reviews are free text and are not limited in length.

The social network is combined with the movie data on the movie information pages within the site (see figure 4). In the next section, we describe how the two are combined.

3.2 Using Trust in FilmTrust

On the movie information pages, we use trust to create predictive movie ratings, that is, we offer a rating that we believe will best match the user's opinion. This is a common task undertaken by collaborative filtering algorithms. While those systems use similarity measures between users to generate a rating, we rely on trust. The basic premise of our algorithm is to compute a weighted average of the ratings assigned to a film, using trust as a weight. If the user does not have a direct trust rating for a person who has rated the movie, we compute an inferred trust value using the TidalTrust algorithm.

3.2.1 TidalTrust: An Algorithm for Inferring Trust

We use trust as a weight in our system, but usually we do not know how much to trust a random person in the system. When two individuals know each other, they can assess the trustworthiness of one another. Two people who are not directly connected do not have a foundation for knowing about trust directly. However, the paths connecting them in the network contain information that can be used to infer how much they may trust one another. Here, we introduce an overview of our algorithm for inferring trust in social networks.

For example, consider that Alice trusts Bob, and Bob trust Charlie. Although Alice does not know Charlie, she knows and trusts Bob who, in turn, has information about how trustworthy he believes Charlie is. Alice can use information from Bob and her own knowledge about Bob's trustworthiness to infer how much she may trust Charlie. This is illustrated in Figure 3.

Our algorithm looks at the trust values along paths connecting the source and sink to compute a recommendation to the source about how much to trust the sink. When making this computation, several features of the network and paths must be considered to produce the most accurate results. In this section, we describe how path length and trust values on paths affect the computations, and how these features were incorporated into our algorithm.

TidalTrust is a modified breadth-first search. The source's inferred trust rating for the sink ($t_{source,sink}$) is a weighted average of the source's neighbors' ratings of the sink (see Forumula 1). The source node begins a search for the sink. It polls each of its neighbors to obtain their rating of the sink. If the neighbor has a direct rating of the sink, that value is returned. If the neighbor does not have a direct rating for the sink, the neighbor queries all of its neighbors for their ratings, computes the weighted average as shown in Formula 1, and returns the result. Each neighbor repeats this process, keeping track of the current depth from the source. Each node will also keep track of the strength of the path to it, computed as the minimum of the source's rating of the node and the node's rating of its neighbor. The neighbor records the maximum strength path leading to it. Once a path is found from the source to the sink, a depth limit is set . Since the search is proceeding in a Breadth First Search fashion, the first path found will be at the minimum depth. The search will continue to find any other paths at the minimum depth. Once this search is complete, the trust threshold (max) is established by taking the maximum of the trust paths leading to the sink. With the max value established, each node completes the calculations of a weighted average by taking information from nodes that they have rated at or above the max threshold. Those values are passed back to the neighbors who queried for them, until the final result is computed at the source.

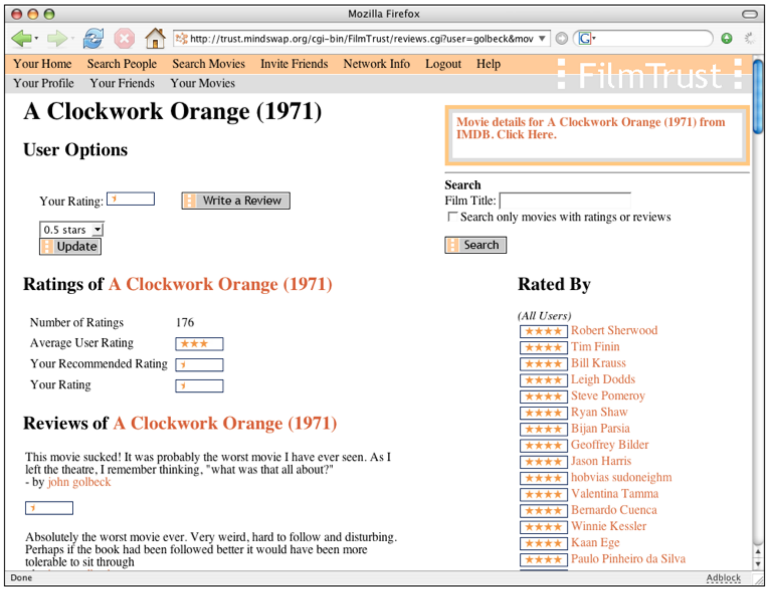
$$t_{is} = \frac{\sum\limits_{j \in adj(j) \ | \ t_{ij} \ \geq \ max} t_{ij} t_{js}}{\sum\limits_{j \in adj(j) \ | \ t_{ij} \ \geq \ max} t_{ij}} \tag{1}$$

On social networks where we have tested this algorithm, it is accurate within approximately 10-15%. A more thorough analysis of the accuracy and implications of error can be found in (Golbeck, 2005).

3.2.2 Creating Predictive Movie Ratings

One of the features of the FilmTrust site that uses the social network is the "Recommended Rating" feature. As figure 4 shows, users will see this in addition to the average rating for a movie.

The "Recommended Rating" is personalized using the trust values for the people who have rated the film (the raters). First, the system searches for raters that the source knows directly. If there are no direct connections from the user to any raters, the system moves one step out to find connections from the user to raters of path length 2. This process repeats until a path is found. The opinion of all raters at that depth are considered. Then, using TidalTrust, the trust value is calculated for each rater at the given depth. Once every rater has been given an inferred trust value, only the ones with the highest ratings will be selected; this is done by simply finding the maximum trust value calculated for each of the raters at the selected depth, and choosing all of the raters for which that maximum value was calculated. Finally, once the raters have been selected, their ratings for the movie (in



A Clockwork Orange (1971)

User Options

Your Rating: ★

0.5 stars

Movie details for A Clockwork Orange (1971) from IMDB. Click Here.

Search

Film Title:

□ Search only movies with ratings or reviews

Ratings of A Clockwork Orange (1971)

Number of Ratings	176
Average User Rating	★★★
Your Recommended Rating	★
Your Rating	★

Reviews of A Clockwork Orange (1971)

This movie sucked! It was probably the worst movie I have ever seen. As I left the theatre, I remember thinking, "what was that all about?"
- by john golbeck

★

Absolutely the worst movie ever. Very weird, hard to follow and disturbing. Perhaps if the book had been followed better it would have been more tolerable to sit through

Rated By

(All Users)

★★★★	Robert Sherwood
★★★★	Tim Finin
★★★★	Bill Krauss
★★★★	Leigh Dodds
★★★★	Steve Pomeroy
★★★★	Ryan Shaw
★★★★	Bijan Parsia
★★★★	Geoffrey Bilder
★★★★	Jason Harris
★★★★	hobvias sudoneighm
★★★★	Valentina Tamma
★★★★	Bernardo Cuenca
★★★★	Winnie Kessler
★★★★	Kaan Ege
★★★★	Paulo Pinheiro da Silva

Figure 4: The movie page on FilmTrust for the film “A Clockwork Orange”.

number of stars) are averaged. For the set of selected nodes S, the recommended rating r from node s to movie m is the average of the movie ratings from nodes in S weighted by the trust value t from s to each node:

$$r_{sm} = \frac{\displaystyle\sum_{j \in S} t_{si} r_{sm}}{\displaystyle\sum_{i \in S} t_{si}} \tag{2}$$

This average is rounded to the nearest half-star, and that value becomes the “Recommended Rating” that is personalized for each user.

As a simple example, consider the following:

- Alice trusts Bob 9
- Alice trusts Chuck 3
- Bob rates the movie “Jaws” with 4 stars
- Chuck rates the movie “Jaws” with 2 stars

Then Alice’s recommended rating for “Jaws” is calculated as follows:

$$\frac{t_{Alice \to Bob} * r_{Bob \to Jaws} + t_{Alice \to Chuck} * r_{Chuck \to Jaws}}{t_{Alice \to Bob} + t_{Alice \to Chuck}} = \frac{9 * 4 + 3 * 2}{9 + 3} = \frac{42}{12} = 3.5 \tag{3}$$

3.3 FilmTrust Evaluation

For each movie the user has rated, the recommended rating can be compared to the actual rating that the user assigned. In this analysis, we also compare the user's rating with the average rating for the movie, and with a recommended rating generated by an automatic collaborative filtering (ACF) algorithm. There are many ACF algorithms, and one that has been well tested, and which is used here, is the classic user-to-user nearest neighbor prediction algorithm based on Pearson Correlation [5]. If the trust-based method of calculating ratings is best, the difference between the personalized rating and the user's actual rating should be significantly smaller than the difference between the actual rating and the average rating.

On first analysis, it did not appear that that the personalized ratings from trust had any benefit over the average. The difference between the user's rating and the recommended rating (call this difference δ_r) was not statistically different than the difference between the user's actual rating and the average rating (call this δ_a). The difference between a user's actual rating of a film and the ACF calculated rating (δ_{cf}) also was not better than δ_a in the general case. A close look at the data suggested why. Most of the time, the majority of users actual ratings are close to the average. This is most likely due to the fact that the users in the FilmTrust system had all rated the AFI Top 50 movies, which received disproportionately high ratings. A random sampling of movies showed that about 50

However, the point of the recommended rating is more to provide useful information to people who disagree with the average. In those cases, the personalized rating should give the user a better recommendation, because we expect the people they trust will have tastes similar to their own [10].

To see this effect, δ_a, δ_{cf}, and δ_r were calculated with various minimum thresholds on the δ_a value. If the recommended ratings do not offer a benefit over the average rating, the δ_r values will increase at the same rate the δ_a values do. The experiment was conducted by limiting δ_a in increments of 0.5. The first set of comparisons was taken with no threshold, where the difference between δ_a and δ_r was not significant. As the minimum δ_a value was raised it selected a smaller group of user-film pairs where the users made ratings that differed increasingly with the average. Obviously, we expect the average δ_a value will increase by about 0.5 at each increment, and that it will be somewhat higher than the minimum threshold. The real question is how the δ_r will be impacted. If it increases at the same rate, then the recommended ratings do not offer much benefit over the simple average. If it increases at a slower rate, that means that, as the user strays from the average, the recommended rating more closely reflects their opinions. Figure 5 illustrates the results of these comparisons.

Notice that the δ_a value increases about as expected. The δ_r, however, is clearly increasing at a slower rate than δ_a. At each step, as the lower threshold for δ_a is increased by 0.5, δ_r increases by an average of less than 0.1. A two-tailed t-test shows that at each step where the minimum δ_a threshold is greater than or equal to 0.5, the recommended rating is significantly closer to the actual rating than the average rating is, with $p < 0.01$. For about 25% of the ratings assigned, $\delta_a < 0.5$, and the user's ratings are about the same as the mean. For the other 75% of the ratings, $\delta_a > 0.5$, and the recommended rating significantly outperforms the average. As is shown in Figure 5, δcf closely follows δa. For $\delta_a < 1$, there

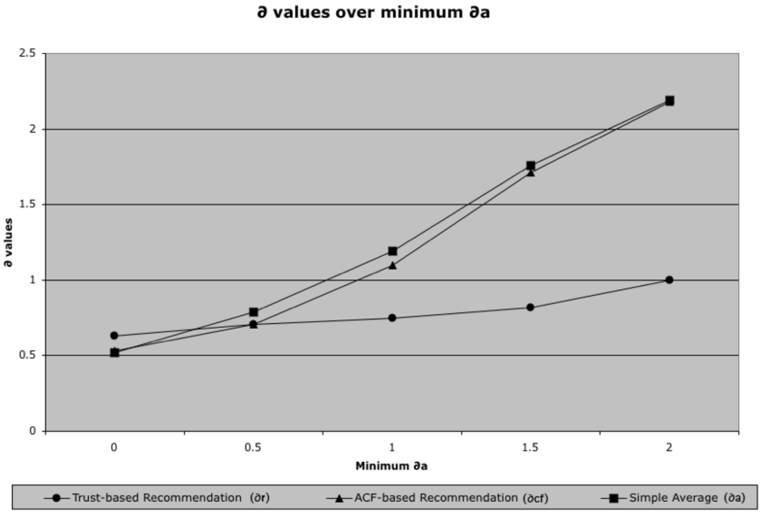


Figure 5: Error rates from different predictive recommendation methods in FilmTrust. Using the simple average, a collaborative filtering algorithm, and a trust-based method all perform equally overall. However, as the threshold for the minimum difference between the user's rating and the average increases, the trust algorithm significantly outperforms both other methods.

was no significant difference between the accuracy of the ACF ratings and the trust-based recommended rating. However, when the gap between the actual rating and the average increases, for $\delta_a >= 1$, the trust-based recommendation outperforms the ACF as well as the average, with p¡0.01. Because the ACF algorithm is only capturing overall correlation, it is tracking the average because most users' ratings are close to the average.

Figure 4 illustrates one of the examples where the recommended value reflects the user's tastes. “A Clockwork Orange” is one of the films in the database that has a strong collective of users who hated the movie, even though the average rating was 3 stars and many users gave it a full 4-star rating. For the user shown, $\delta_a = 2.5$ a very high value while the recommended rating exactly matches the user's low rating of 0.5 stars. These are precisely the type of cases that the recommended rating is designed to address.

3.4 Motivation

These results suggest that when users assign trust ratings, they are capturing more than simply how similar they are to others. If trust were only a reflection of overall similarity, we would expect closer alignment of the performance from ACF algorithms and trust algorithms. What precisely is expressed in trust beyond overall similarity? That question motivates the experimental work in the rest of this paper.

4. The Experiment

In order to understand how similarity on different profile attributes affect the trust people assign to one another, we conducted a survey-based experiment. Subjects began by rating a large number of movies. We then generated profiles of hypothetical users that included ten to twenty films. In a profile, the subject was shown the hypothetical user's rating for each film, along with the subject's own rating and the average rating. Subjects were asked to say how much they would trust this hypothetical user based on the profile. Profiles varied when and how much the hypothetical person agreed with the user.

4.1 Subject Background

Fifty-nine subjects participated in this study. They ranged in age from 20 to 52, with an average age of 32 (standard deviation of 8.5 years). Education levels were mixed among subjects who reported it: six subjects were current undergraduates, eleven had bachelor's degrees, twenty-three had master's degrees, and eleven had Ph.Ds.

On average, subjects reported watching movies about once a week, and looking at movie related media and websites every week or two.

Further details about subjects' movie rating patterns are contained in section 4.2.2.

4.2 Stage One: Learning Subjects' Opinions

4.2.1 Compiling a List of Movies

Subjects began the study by rating a large collection of movies. In choosing movies to include as part of this study, there were several factors we considered important to create a fair coverage:

1. Movies most subjects would have seen - We wanted our list of movies to contain films that had reached a wide audience, so that subjects could to rate a large number of films. To ensure that many popular movies were in the mix, we included the 100 worldwide top grossing films of all time[1].

2. Cover a broad spectrum of genres - For subjects who had interests mostly in specific genres, we took the top 10 rated movies from each genre as listed in the Internet Movie Database (IMDB): Action, Adventure, Animation, Family, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Independent, Musical, Mystery, Romance, Science Fiction, Thriller, War, and Western.

3. Include bad movies - The nature of the first two considerations necessarily means that our movies would be skewed toward the good end. To be sure that plenty of bad films were on the list, we looked at the IMDB 100 worst rated movies. We took all the films on that list that had at least 1,000 ratings to ensure that these were bad movies that our subjects may have seen. Once the movies with under 1,000 ratings were removed, 65 bad movies were left to be included in our set.

After making these selections, some films were duplicated (e.g. Lord of the Rings: Return of the King was one of the top grossing movies as well as appearing in the top ten lists of the action, adventure, and fantasy genres). With duplicates removed, we were left with a total of 283 films. This made up the core set of movies that we used in the study.

In stage one of the experiment, we presented users with a list of these 283 films and asked them to assign a rating to each movie that they had seen and about which they had an opinion. Ratings were on a 1-10 scale (1 for bad movies, and 10 for good ones). This rating system matched the Internet Movie Database scale, which allowed us to use the IMDB average ratings in our analysis.

4.2.2 Movie Rating Statistics

Subjects rated an average of 136.4 movies (standard deviation 51.3). Ratings for movies were normally distributed around a mean of 6.3 (standard deviation 2.4) on a scale of 1-10. This was slightly skewed to the high end because many of our movies were pulled from top-rated lists. The distribution of ratings is shown in in figure 6.

Most subjects ratings were distributed around the mean, however there were some outliers. For example, one subject assigned a rating of 1 to 81 of the 183 movies he or she rated while the average subject (excluding this outlier) only rated 3.3 movies with a 1. This subject alone was responsible for over 25% of all the ratings of 1.

4.3 Stage 2: Assigning Trust to Profiles

To measure how trust correlated to similarity on certain profile features, we generated profiles representing the preferences of hypothetical users and asked subjects to rate how much they trusted those people.

[1] http://imdb.com/boxoffice/alltimegross?region=world-wide

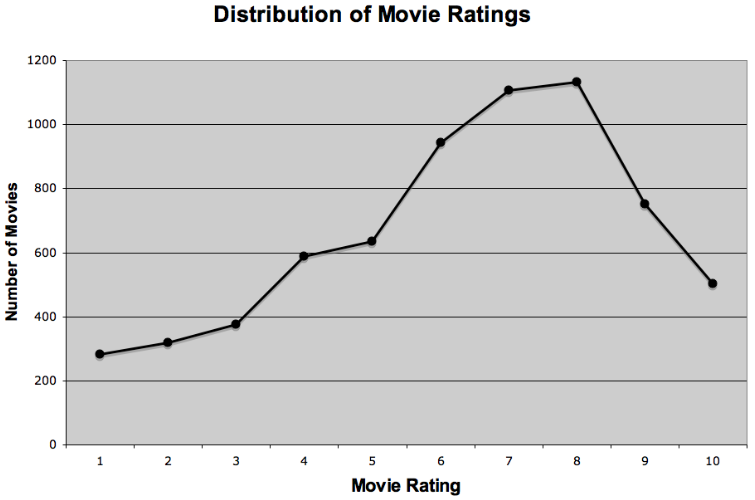


Figure 6: Distribution of Movie Ratings Assigned by Subjects

4.3.1 Generating Profiles

After rating the films, users were presented with a series of profiles. This work was conducted in two phases. These were designed to test different conditions.

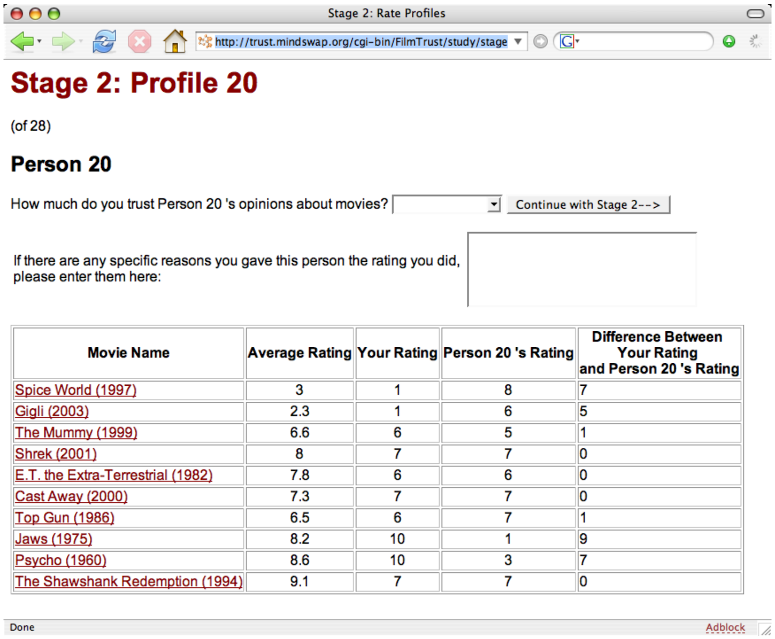
Phase 1 In the first phase, subjects were presented with 28 profiles ($P_1 - P_{28}$). Each profile contained ten to sixteen films chosen from the set of movies the subject rated. For each film in the profile, subjects were shown the average rating for each film, their own rating, and a rating generated to represent the opinion of a hypothetical user.

Call the subject's rating of movie i r_{si}, the profile's rating of the same movie r_{pi}, and the average rating r_{ai}. Call the absolute difference between any two ratings Δ. There are three categories(C_1 - C_3) where we initially expected to see a correspondence of trust and specific profile features:

C_1: Extreme Ratings ($r_{si} \leq 2$ or $r_{si} \geq 9$) - If the subject rated a movie 1, 2, 9, or 10, we considered the rating to be on the extreme ends of the scale.

C_2: Far from average ($\Delta_{r_{si},r_{ai}} \geq 4$)- Among our subjects, the average difference between the subject's rating and the average was 1.82, with a standard deviation of 1.78. Since the data was normally distributed, approximately 95% of the ratings were within ±4 of the average. Thus, we looked for a difference ≥ 4 to qualify a rating as largely different from the average.

C_3: Extreme ratings that are farr from the average ($\Delta_{r_{si},r_{ai}} \geq 4$ and either $r_{si} \leq 3$ or $r_{si} \geq 9$) - This occurs when the difference between the subject's rating and the average is ≥ 4 and the subject's rating is 1, 2, 9, or 10.



Stage 2: Profile 20

(of 28)

Person 20

How much do you trust Person 20 's opinions about movies? Continue with Stage 2-->

If there are any specific reasons you gave this person the rating you did, please enter them here:

Movie Name	Average Rating	Your Rating	Person 20 's Rating	Difference Between Your Rating and Person 20 's Rating
Spice World (1997)	3	1	8	7
Gigli (2003)	2.3	1	6	5
The Mummy (1999)	6.6	6	5	1
Shrek (2001)	8	7	7	0
E.T. the Extra-Terrestrial (1982)	7.8	6	6	0
Cast Away (2000)	7.3	7	7	0
Top Gun (1986)	6.5	6	7	1
Jaws (1975)	8.2	10	1	9
Psycho (1960)	8.6	10	3	7
The Shawshank Redemption (1994)	9.1	7	7	0

Figure 7: A sample profile from the experiment.

Each profile, P_j, was generated by choosing a set of movies, M_k from the movies rated by the subject. Four to six movies came from a category C_m and six to ten movies from its complement. For example, a profile could comprise four movies where the subject had given extreme ratings, and six movies where the subject's ratings were not extreme. Once the movies were selected, we generated the profile rating for each film.

We tested for small, medium, and large differences between the subject's rating and the profile's rating by randomly generating a Δ in the correct range:

1. Small variations: $0 \leq \Delta_{r_{si},r_{pi}} \leq 1.5$
2. Medium variations: $0 \leq \Delta_{r_{si},r_{pi}} \leq 6.5$
3. Large variations: $0 \leq \Delta_{r_{si},r_{pi}} \leq 9.5$

For profile P_j, r_{pi} is created for each of the four to six movies selected from the category C_m, and for each of the six to ten movies chosen from it's complement, $\overline{C_m}$. For example, profile P_3 was generated with small differences on movies with extreme ratings(C_1), and large differences on movies with non-extreme ratings($\overline{C_1}$). The number of profiles with small (s), medium(m), and large(l) differences among the categories is shown in table 1.

Subjects were presented with a total of 28 profiles and asked to rate how much they trusted the person represented by the profile. Trust was assigned on a scale of 1 (low trust) to 10 (high trust). There was also space to comment on any specific reasons for the rating.

Table 1: This table shows the number of profiles presented to each subject in each category with the shown small, medium, and large differences. For example, (s,l) on $(C_1, \overline{C_1})$ would indicate a small difference between the profile and user ratings on movies from category C_1, and a large difference between the profile and user ratings on movies in from the complement of category C_1. (Note: No (s,s) profile was used for C_3 because the small differences were addressed in C_1 and C_2, and C_3 is just the intersection of C_1 and C_2.)

	$(C_1, \overline{C_1})$	$(C_2, \overline{C_2})$	$(C_3, \overline{C_3})$
(s,s)	2	2	0
(s,m)	4	4	1
(s,l)	2	2	1
(m,s)	2	2	1
(l,s)	2	2	1

Table 2: For each pre-defined set of rating differences (Ψ_i), ten individual differences (ψ_j) are enumerated.

	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	ψ_8	ψ_9	ψ_{10}
Ψ_1	1	1	1	1	0	0	0	0	0	0
Ψ_2	4	4	4	4	1	0	1	1	0	0
Ψ_3	6	6	6	6	1	0	1	1	0	0

Phase 2 Just as in Phase 1, the second phase of the study was conducted by presenting profiles with generated ratings to the subjects, who were asked to rate how much they trusted the hypothetical user represented by the profile. However, instead of generating differences between the subject's rating and the profile rating that fell within a range, we used fixed sets of rating differences and chose to distribute them in different ways.

The variation among profiles in the first phase was useful because it provided data on many different features of profile similarity. To test specific hypotheses, a more controlled experiment was required. In phase two, each profile consisted of ten movies. Exactly four were chosen from a category, and six from its complement.

Each set of movies in the second phase was chosen from the user's set of rated films based on the categories in table 2. We then established three pre-defined sets of rating differences ($\Psi_1 - \Psi_3$). Each set contained ten values, $\psi_1 - \psi_{10}$. The Ψ_n values were applied to the movie ratings in each profile, such that for movie j and Ψ_n, $\Delta_{r_{sj},r_{pj}} = \Psi_n$. This insured that the average rating, the variance, the distribution of ratings, and the largest $\Delta_{r_{si},r_{pi}}$ was the same among all profiles using a given Ψ. Thus, for a fixed Ψ, the only variation among profiles is which movies the Ψ_n are applied to, and this allows us to test if variation on certain categories of films is more important than others.

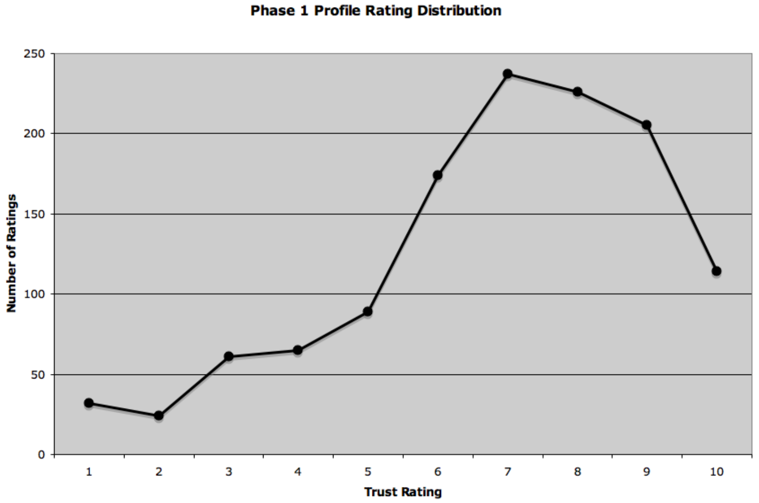


Figure 8: Distribution of Profile Trust Ratings Assigned by Subjects, Phase 1

Once the films were selected, the user's rating for each movie was paired with a ψ_j to produce the profile rating.

Let $\psi_{high} = \{\psi_1, \psi_2, \psi_3, \psi_4\}$ and $\psi_{low} = \{\psi_7, \psi_8, \psi_9, \psi_{10}\}$. For each $(C_m, \overline{C_m})$ we selected a set of ten movies, M_k. Four movies $(m_1 - m_4)$ were chosen from C_m and six $(m_5 - m_{10})$ from $\overline{C_m}$. Profile ratings were generated by pairing $m_1 - m_4$ with either ψ_{low} or ψ_{high} and $m_5 - m_{10}$ with the remaining ψ values. For each $(C, \Psi, \{high, low\})$ combination, subjects were shown three profiles, for a total of 54 profiles to be rated.

4.3.2 Profile Rating Statistics

Figures 8 and 9 show the overall distribution of trust ratings assigned to profiles in Phases 1 and 2 of the study. In Phase 1, the ratings have a normal distribution, skewed to the high end with an average of 6.90. This distribution and mean are expected based on the way profiles were generated; the Δ for each movie was randomly generated within a given range, and more profiles had small differences than big ones, leading to the high average trust rating. In Phase 2, the distribution of ratings is less obviously normal, which is explained by the use of the Ψ distributions to create profiles. The average trust rating was also lower in Phase 2, at 5.95.

5. Experimental Results

In phase 1, the profiles had a wide range of characteristics. We measured agreement on a profile P_j by the average $\Delta_{r_{si}, r_{pi}}$ for all movies in the profile. Call the average difference on any set of movies Θ, and let the overall average difference on all movies in P_j be $\overline{\Theta}_j$.

We are interested in identifying which features of profiles, if any, correlate more strongly with trust than overall agreement. For example, does agreement on movies where the user

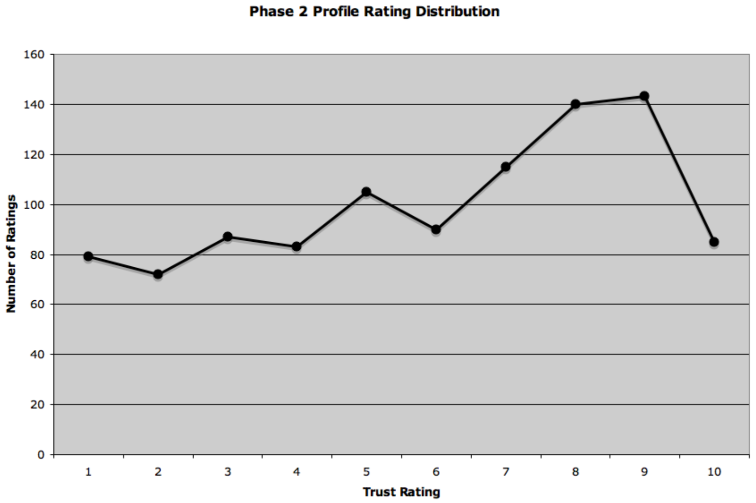


Figure 9: Distribution of Profile Trust Ratings Assigned by Subjects, Phase 2

gave an extreme rating matter more than good overall agreement? Does the user's difference from average ratings matter more? In this section we analyze the data to identify how these features correlate with trust.

5.1 Correlation of Trust and Overall User Similarity

As mentioned in the previous work, (Ziegler & Golbeck, 2006) established a relationship between trust and user similarity. We began by testing the correlation of trust with overall similarity to show that same result on our data. Correlation was measured by computing the Pearson correlation coefficient. We first checked the correlation of trust assigned to a profile P_j (call this t_j), and $\overline{\Theta}_j$. That correlation was -0.65, indicating a strong negative correlation; that is, as the average absolute difference increases, trust decreases.

To gain more insight into the correlation of trust and overall similarity, we looked at intervals of $\overline{\Theta}$. We took all of the profiles with $0 \leq \overline{\Theta} < 1$ and grouped them together. We did the same for $1 \leq \overline{\Theta} < 2$, $2 \leq \overline{\Theta} < 3$, etc. Call these groupings $\overline{\Theta}^{[a,b)}$. With this grouping, we analyzed the trust value assigned to each profile in the group, and thus we could track how trust changed as $\overline{\Theta}$ varied. Table 3 shows the average trust value for each range of $\overline{\Theta}$ values. An ANOVA test on this data showed that the trust values were significantly different based on $\overline{\Theta}$ for $p < 0.001$. Using a standard 2-tailed t-test, we compared the trust values for each consecutive range of $\overline{\Theta}$ values (e.g. we compared the trust values for $\overline{\Theta}^{[0,1)}$ to those for $\overline{\Theta}^{[1,2)}$, $\overline{\Theta}^{[1,2)}$ to $\overline{\Theta}^{[2,3)}$, and so on). Our results showed a significant difference between each consecutive pair from $\overline{\Theta}^{[0,1]}$ to $\overline{\Theta}^{[5,6)}$ with $p < 0.05$. Beginning with the comparisons of $\overline{\Theta}^{[5,6)}$ to $\overline{\Theta}^{[6,7)}$, the differences were not statistically significant because there were very few datapoints. There were no profiles with $\overline{\Theta} > 8$, and only four profiles with $\overline{\Theta}^{[7,8)}$. To reach such high overall average differences, the user would need to have most of their ratings

Table 3: Average trust values for ranges of $\overline{\Theta}$. Note that there were no profiles with $\overline{\Theta} \geq 8$. That would require all of the users ratings to be very high or very low to allow profile ratings to have such a large difference on the majority of movies. For similar reasons, there were less than ten movies in the range [6,7) and [7,8).

$\overline{\Theta}$ **Range**	**Average Trust Rating**
[0,1)	8.01
[1,2)	6.44
[2,3)	5.27
[3,4)	3.96
[4,5)	2.72
[5,6)	1.94
[6,7)	2.09
[7,8)	2.25

either very high or very low to allow for the possibility of such a big difference with most of the profile ratings.

These results support the previous research that trust correlates with user similarity. We see significant increases in trust values as the overall agreement increases.

5.2 Phase 1 Results: The Maximum Difference Effect

With the data in Phase 1, we began by looking at agreement on movies where the user had assigned extreme ratings ($r_{si} \leq 2$ or $r_{si} \geq 9$). It could be that extreme ratings indicate that the user is more passionate about the movie, and thus it would matter more to the user that the profile agrees on these films over those where the rating is in the middle range. Our first experiment to check this was to hold overall agreement ($\overline{\Theta}$) constant and then check the correlation between trust and agreement on extremes. With $\overline{\Theta}$ held constant, if agreement on extremes increases, then agreement on non-extremes ($3 \leq r_{si} \leq 8$) must decrease.

Holding $\overline{\Theta}$ constant, as agreement on extremes increased, trust increased. This initially would indicate that trust is tied, in part, to agreement on extremes. However, when we held $\overline{\Theta}$ constant and increased agreement on *non-extremes* (and thus agreement on extremes decreased), trust *also increased*. These results indicated that some factor other than extremes or non-extremes was involved and wat affecting the trust values.

Consider what happens when the average difference between the user's ratings and the profile's ratings increases. There are two possibilities - the number of movies where there is disagreement must increase, or the size of the $\Delta_{r_{si},r_{pi}}$ must increase. If the overall average difference is heald constant, and agreement increases in one category, that means agreement must *decrease* over the rest of the movies. Thus, on the set of movies where agreement decreases, the standard deviation among those movies may increase, or the differences between the user and profile ratings on movies will grow larger. It is possible that these changes affect trust more strongly than extreme or non-extreme values. To test how these factors correlate to trust, there are two tests we conducted: the correlation of trust with the

Table 4: For the two $(\sigma, \overline{\Theta})$ pairs that allowed a statistical analysis, this table shows the ranges of ∇ and the corresponding average trust value.

$(\sigma, \overline{\Theta})$=([0,1), [0,1))				
∇ Value	0-1	1-2	2-3	3-4
Average Trust Value	9.11	8.58	8.16	7.52

$(\sigma, \overline{\Theta})$=([1,2), [1,2))				
∇ Value	2-3	3-4	4-5	5-6
Average Trust Value	7.30	6.61	6.22	5.49

standard deviation (σ) on a profile, and the correlation of trust with the maximum $\Delta_{r_{si},r_{pi}}$ on a given profile. Call this maximum $\Delta_{r_{si},r_{pi}}$ ∇.

As ∇ increases, σ will also increase. To see the effect of ∇ independent of σ, we looked at the correlation of ∇ and trust with σ and $\overline{\Theta}$ held in fixed ranges. We used integer ranges for σ and $\overline{\Theta}$ (e.g. [0,1), [1,2), etc.). This created 100 smaller datasets where we could examine the effect of ∇ on trust (e.g. one of these data sets is σ in the range (0,1] and $\overline{\Theta}$ in the range (2,3]). Within each $(\sigma, \overline{\Theta})$ pair, trust values were grouped by ∇ in integer ranges ($0 \leq \nabla < 1, 1 \leq \nabla < 2$, ...). If ∇ is a factor in how users assign trust, we expect to see significant changes in trust when ∇ increases while σ and $\overline{\Theta}$ are held constant.

There were approximately 1,200 trust values assigned in Phase 1 of the study. Because our data was broken up into 100 smaller datasets, which in turn divide the data into ten groups, very few $(\sigma, \overline{\Theta})$ pairs had enough data to allow for a statistical analysis. We expect to have very little data in the higher ranges for σ and $\overline{\Theta}$, because profiles with such large differences on every movie were not considered. Similarly, with a low $\overline{\Theta}$, it is not mathematically possible to have high values of $sigma$. In fact, approximately 80% of our data came from profiles where $\sigma < 2$ and $\overline{\Theta} < 2$.

We were able to perform a statistical analysis of the ∇ - trust relationship on two $(\sigma, \overline{\Theta})$ pairs:

1. $(\sigma, \overline{\Theta})$ = ([0,1), [0,1)) : 416 trust ratings, about 35% of the total number of trust ratings made in phase 1.

2. $(\sigma, \overline{\Theta})$= ([1,2), [1,2)) : 355 trust ratings, about 30% of the total number of trust ratings made in phase 1.

In both of these cases, which together covered approximately 65% of all the trust ratings given in phase 1, we see a very clear relationship between ∇ and trust.

Table 4 shows the average trust values as ∇ increases. In both cases, we can see the average trust rating decreases as ∇ increases. An ANOVA showed that there were significant differences within the population for $p < 0.01$ when $(\sigma, \overline{\Theta}) = ([0, 1), [0, 1))$ and also when $(\sigma, \overline{\Theta}) = ([1, 2), [1, 2))$. Pairwise comparisons using a 2-tailed t-test were made for each population $i \leq \nabla < i+1$ and $i+1 \leq \nabla < i+2$. For the populations shown in Table 4, each pairwise comparison was statistically significant for $p < 0.05$. This result suggests that ∇, the maximum difference between the user's rating and profile's rating of a given movie, is a significant factor in how users assign trust. Specifically, as the single largest difference increases, trust decreases.

5.3 Phase 2 Results: Extreme Ratings

In phase 1, we could not see if extreme ratings and trust had a relationship because the profiles had other factors that were impacting trust whereas, in phase 2, the profiles were much more controlled. Profiles used one of three sets of rating differences (Ψ_n), and the ψ values were distributed among a set of ten movies, selected for certain features. For example, if a profile were using Ψ_2 and the ψ_{high} values were being assigned to movies the user gave extreme ratings, a sample profile might look like what is shown in figure 5.

Thus, for a given Ψ_n, the only variable was to which movies the ψ values were assigned. $\overline{\Theta}$, ∇, σ, and other factors were all constant.

Our first hypothesis was that when the larger ψ values in Ψ_n were applied to movies where the user gave an extreme rating (C_1), that users would have lower trust than when the psi_{high} values were used on $\overline{C}_1$ movies.

For each Ψ_n, we compared the average trust rating assigned to a profile when ψ_{high} were used on movies with extreme ratings vs. when ψ_{high} were applied to movies with non-extreme ratings. For Ψ_1, there was no statistically significant difference. We believe this is because users consider a maximum difference of 1 on a 1-10 scale to be trivial. For Ψ_2 and Ψ_3, the average trust ratings were significantly lower ($p < 0.05$) for profiles where movies with extreme ratings had the large differences.

Table 6 shows the average trust values assigned to profiles for each Ψ_n based on which films received the ψ_{high}. When ψ_{high} were assigned to movies with extreme ratings, the average trust value for the profile averaged over 0.5 lower than when the movies with extreme ratings used ψ_{low}.

From this, we are able to conclude that agreement on movies to which the subject has assigned extreme ratings (Θ_χ) affects trust, independently of overall agreement. When agreement is closer on these movies, trust is higher.

5.4 Phase 2: Trust Predicting Trust

As discussed in the previous work, there are factors specific to each person and each relationship that affect trust independently of similarity. While most of that information is not available in the types of systems we are studying, we *do* have the user's own trust ratings, which provide some insight.

In a real system, like FilmTrust, users may make choices about who their friends are that impact the trust ratings they assign. For example, someone who only connects to and rates their closest friends may assign trust values with higher average than someone who rates a large group of acquaintances. Because of unknowable factors like this, it is difficult to analyze a real system and determine why some users assign, on average, higher trust values than others.

Table 5: A sample Phase 2 profile using Ψ_2 with ψ_{high} on movies with extreme ratings.

Movie	**User Rating**	**Profile Rating**	ψ
Movie 1	9	5	4
Movie 2	10	6	4
Movie 3	1	5	4
Movie 4	2	6	4
Movie 5	6	7	1
Movie 6	4	4	0
Movie 7	8	7	1
Movie 8	3	2	1
Movie 9	5	5	0
Movie 10	7	7	0

Table 6: Average trust ratings in phase 2 for each Ψ_n with ψ_{high} distributed to extremes or non-extremes.

	ψ_{high} Assigned To	
	Extremes	Non-Extremes
Ψ_1	8.62	8.72
Ψ_2	5.23	5.83
Ψ_3	2.65	3.34

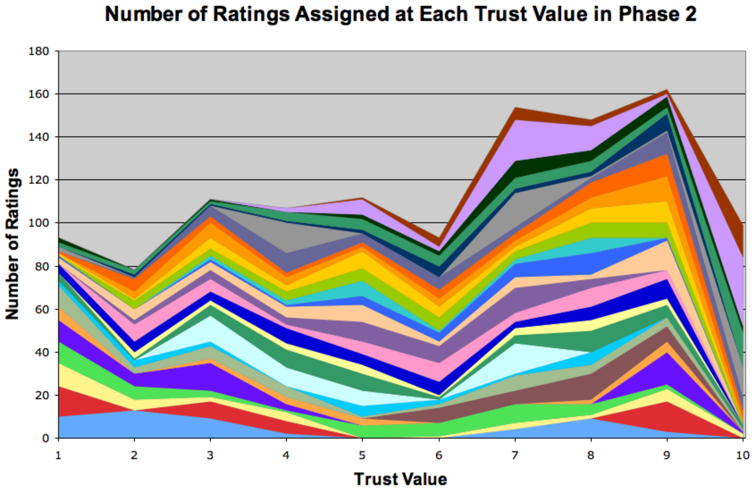


Figure 10: Distribution of Trust Ratings assigned by 29 subjects in Phase 2

In Phase 2 of our experiment, though, all the subjects were rating the same population of hypothetical users. The profiles were carefully controlled, so the profiles rated by each subject all varied from the user in the same way. That is, Profile i in Phase 2 contained ten movies selected on the same criteria for all users (e.g. four movies where the user gave extreme ratings and six where the user gave non-extreme ratings) where the profile's rating differed from the user's rating by the same amount on each type of film. Thus, we can look the average trust value assigned by user s (call this τ_s) to estimate how "trusting" the user is.

Figure 10 shows the ratings assigned by 29 subjects in Phase 2. As is shown, the distribution of ratings varies greatly from user to user. The average trust rating assigned by subjects to profiles in phase 2 ranged from as low as 3.73 up to 8.48. Note that the top two layers in our stack, representing the most trusting users, have no ratings below 4, while the bottom two layers represent users who have at least 2/3 of their ratings at 4 or lower. With that range of τ among users rating sets profiles that varied identically for each subject, it indicates that some subjects are inclined to assign higher trust values that others. It could be that some subjects are more trusting than others, or that subjects interpret the scale of trust values differently. Whatever the reason, the average trust value has a wide range.

The τ for a given user informs what trust values we expect them to assign in the future. If a user is consistently giving low trust ratings we can anticipate that their future ratings will also be lower. Since we are only considering the trust ratings the user has assigned, there are no issues of privacy or access control to worry about; the information is simply part of the user's profile. We include a τ in our analysis in section 6.

5.5 Difference from Average

In both Phase 1 and Phase 2 of the experiments, we tested to see if agreement and disagreement in C_2, where the user's rating was far from the average rating, affected trust. When considering this, we do not want to conflate these results with those of extreme ratings. Overall, 28% of the movies in the system were given extreme ratings (rated 1,2, 9, or 10). In the set of movies where the user's rating was far from the average (a difference of at least 4), 50.4% of the movies had extreme ratings. Since the proportion of extreme ratings was significantly higher in this group, it would be possible for the impact of extreme ratings to make it appear as though difference from average was also significant. Thus, we controlled for extreme ratings in our analysis.

With the data from Phase 1, we tested to see if large differences on movies in C_2 caused trust to change significantly from when there were only small differences on movies in C_2. We were unable to find any statistically significant results overall, nor when we eliminated movies with extreme ratings from the pool. We performed the same analysis in Phase 2. Again, overall and when we controlled for extreme ratings, there was no significant impact visible in our results.

C_3 considered both extreme ratings *and* difference from average. An analysis here did not indicate any results beyond what we saw in C_1 when looking at extreme ratings only. Since C_2 did not indicate significant differences in trust values, this is not surprising.

It was a somewhat surprising result that difference from average did not affect trust, since the initial results of the FilmTrust study described in section 3 were seen when the user's opinion grew apart from the average. While that analysis was checking the accuracy of recommended ratings, not the correlation of trust ratings, we expected to see a parallel result in this study. However, from the large number of users and data we gatherd here, we can only conclude that agreement or disagreement on movies where the user is far from the average does not have a significant impact on user's trust ratings.

5.6 Experimental Conclusions

From the survey experiments conducted here, our results suggest that overall difference ($\overline{\Theta}$), the maximum single difference in movie ratings (∇), difference on movies to which the user assigned extreme ratings (Θ_χ), and the user's average trust value (τ) all are factors in how users assign trust. However, it is difficult from these experiments to know how strongly each of those features affect trust. In the next section, we use these parameters in FilmTrust and show how they can be composed to predict trust more accurately and precisely than using overall agreement alone.

6. Predicting Trust with Profile Similarity

The study results are controlled, and many aspects of trust cannot be accounted for. In a real system, social features independent of similarity will certainly affect the trust ratings people assign to one another. These social features are, in fact, the more studied aspects of trust. A person's propensity to trust in general, the nature of the relationship with the person being trusted, their history outside of the system, and many other factors will impact the trust rating that is assigned, and those factors are largely independent of any

Table 7: Statistics relating actual trust values and predicted trust values derived from overall agreement and equation 4 .

	Overall	Equation 4
Correlation With Actual Trust Value	0.24	0.68
Absolute Mean Error of Computed and Actual Trust Value	1.91	1.22
Standard Deviation of Average Difference	1.95	1.09

profile similarity measures. The question we face is whether or not the aspects of profile similarity that correlate with trust can be used to predict trust in a real system.

To test this, we looked at data taken from the FilmTrust system. When two people are connected in the FilmTrust social network, call the person doing the trusting the *source* and the person being trusted the *sink*. For each ordered pair of users, we collected the trust value, the average trust value assigned by the source (τ), number of movies in common, overall error ($\overline{\Theta}$), overall error on extremes (Θ_{χ}), and the maximum difference (∇). Our goal was to see if a composition of the values could yield a relatively accurate estimate of trust. To ensure we had enough data to work with, we eliminated any pairs with less than 10 movies in common. That left us with 366 user-pairs to analyze.

Unlike the controlled environment of our study where certain variables could be held constant, these measures of error are all interrelated. The Pearson correlation coefficient for $\overline{\Theta}$ and ∇ is 0.42. $\overline{\Theta}$ and Θ_{χ} also had a correlation coefficient of 0.42. Θ_{χ} and ∇ had a high correlation of 0.70. For some pairs, Θ_{χ} did not exist because the sink had not rated any movies given an extreme rating by the source.

Using these $\overline{\Theta}$, Θ_{χ}, and ∇ values, along with the τ for the source to adjust our estimate up or down, we were able to predict trust rather accurately. Equation 4 gives the equation we fine tuned to our dataset. The weights (w values) and thresholds (l) used in the system ($w_{\overline{\Theta},1}, w_{\overline{\Theta},2}, w_{\nabla,1}, w_{\nabla,2}, w_{\Theta_{\chi}}, l, w_{\tau}$) were set to (6.83, 9.03, 1.99, 0.97, 1, 7.1, 0.835).

$$t_{ij} = min\left(10,\ \begin{array}{lr} \lfloor w_{\overline{\Theta},1}\overline{\Theta}_{ij} + w_{\nabla,1}\nabla_{ij} + w_{\Theta_{\chi}}\Theta_{\chi,ij} - w_{\tau}(l-\tau_i)\rceil & \text{if } \Theta_{\chi} \text{ exists} \\ \lfloor w_{\overline{\Theta},2}\overline{\Theta}_{ij} + w_{\nabla,2}\nabla_{ij} - w_{\tau}(l-\tau_i)\rceil & \text{otherwise} \end{array}\right) \tag{4}$$

We computed the accuracy of trust values predicted by formula 4 by taking the absolute mean error of the computed trust value and the known trust rating. This was compared to the accuracy of trust predicted using only overall agreement. As is shown in table 7, our trust estimation using the set of profile similarity measures has an absolute mean error of 1.22, a 36% increase in accuracy over trust prediction using overall similarity alone. A two-tailed t-test shows that these results are statistically significant for $p < 0.001$. The correlation was also quite high, with a Pearson Correlation Coefficient of 0.68 using this formula vs. 0.24 using overall agreement alone.

While formula 4 and the values used in it are fine tuned to the FilmTrust system, the large increase in accuracy afforded by incorporating these profile similarity measures is a good indicator that other systems can incorporate the measures to estimate the trust users have in one another.

These results from FilmTrust are a strong reinforcement to the results obtained in section 5. Using overall error ($\overline{\Theta}$), overall error on extremes (Θ_{χ}), and the maximum difference (∇) along with the user's average trust rating (τ), we significantly improved the correlation, accuracy, and standard deviation of trust predictions from profile information. This supports our previous analysis that $\overline{\Theta}$, Θ_{χ}, ∇ all impact trust, since our optimized formulation included non-zero weights for each parameter. Furthermore, because these parameters are effective in a real system, it shows that our survey results are not merely the product of a tightly controlled experiment, but rather are more generally applicable set of principles regarding trust and profile similarity.

7. Discussion

When working in a system that has trust and allows users to rate items, there are two methods for determining how much a source should trust a sink: predicting trust using the social network, and predicting trust using profile similarity. Using trust values in a social network, trust can be computed by gathering information from friends, who they trust, and how much (as discussed in section 2.3). As has been shown in previous work, this can be effective when the users are connected into a large social network where trust values are accessible. However, in many social networks a large percentage of users are completely isolated from most others, and social network-based techniques are not applicable.

To compute trust in these cases, we only have access to the source's trust ratings, the source's profile information and the sink's profile information. From the experimental survey results and their application in the FilmTrust system, we have shown that trust correlates more strongly with a measure that includes several facets of profile similarity than with simple overall similarity.

These results are useful for several applications:

1. Recommender Systems - Earlier work (Sinha & Swearingen, 2001), (Swearingen & Sinha, 2001) have shown that people prefer recommendations from trusted sources rather than from recommender systems. While there is certainly a social aspect to this, our work presented here shows that trust also captures a more nuanced similarity than what is traditionally used in recommender systems. We believe it is likely that one reason users prefer recommendations from trusted people is because those people share these nuanced preferences. If a recommender system can replicate the facets of similarity that are captured by a trust relationship, it may be possible to produce recommendations more like those from trusted people, that in turn are more appreciated by users. The results from this study can give guidance on how recommender systems might begin incorporating new information.

2. Refining Trust Inference Algorithms - In our literature search, we found no algorithms that use trust relationships and connections in a social network to infer trust *as well as* profile similarity. When profiles and social network information is available, it may be possible to improve the accuracy of the trust inference algorithms by using all of the available information.

3. Trust Estimation for Intelligent Systems - In this work, the only application of trust computed from a social network that we present is FilmTrust, a recommender system. There are other applications of trust. In other work we have discussed the use of trust for priorities in default logics (Katz & Golbeck, 2006), the basis for a trust-based policy system. We are currently also looking at using trust in other policy systems and for belief revision. In these applications, we need an accurate estimate of trust. As discussed before, when there is a strongly connected component, social network-based algorithms can be used to infer a trust rating. However, when a large number of users are not connected, another mechanism is needed to compute trust. We have shown that trust can successfully be estimated using the profile features discovered in our survey, and that can fill in the gaps left network-based trust algorithms.

8. Conclusions and Future Work

Up to now, the research has only shown a correlation between trust and overall similarity. Through the surveys conducted here, we have shown that in addition to overall similarity, there is also a correlation between trust and the largest single difference in ratings, and between trust and the agreement on movies the source has given extreme ratings. We have also shown that some sources tend to assign higher ratings than others when rating a population of sinks that vary from the source in the same way.

We then showed in the FilmTrust system that a composition of these measures can be used to predict trust with a stronger correlation, better accuracy, and less variation than when using overall agreement alone. This shows that the insights obtained in our study can be implemented in a real system to improve the accuracy of trust inferences.

We plan to extend our work in this space by investigating how these results can be incorporated into algorithms and intelligent systems. We are in the early stages of work for creating policy systems using trust. In these systems there is a social network backing, but using policy sets as profiles may also allow an application of these results such that trust could be estimated even when the social network does not exist.

One of our ongoing projects is to investigate how this type of profile analysis can be integrated with the network-based trust algorithms to improve the accuracy of the results. We are also looking at whether these similarity measures can be used with trust ratings as the basis instead of movie ratings. While the structure of social networks suggests this would only be effective for users within a small distance of the user, if finding maximum differences, agreement on extremes, and overall agreement on trust ratings assigned to common friends can improve the accuracy of predicted trust, it would be a worthwhile modification of the algorithm.

9. Acknowledgments

This work, conducted at the Maryland Information and Network Dynamics Laboratory Semantic Web Agents Project, was funded by Fujitsu Laboratories of America – College Park, Lockheed Martin Advanced Technology Laboratory, NTT Corp., Kevric Corp., SAIC, the National Science Foundation, the National Geospatial-Intelligence Agency, DARPA, US Army Research Laboratory, NIST, and other DoD sources.

References

Burgess, E., & Wallin, P. (1943). Homogamy in social characteristics. *American Journal of Sociology*, *2*(49), 109–124.

Byrne, D. (1961). Interpersonal attraction and attitude similarity. *Journal of Abnormal and Social Psychology*, *62*, 713–715.

Byrne, D. (1971). *The Attraction Paradigm.* Academic Press, New York, NY, USA.

Castelfranchi, C., & Falcone, R. (1998). Principles of trust for mas: Cognitive anatomy, social importance, and quantification. *3rd International Conference on Multi Agent Systems.*

Castelfranchi, C., & Falcone, R. (2002). Social trust: A cognitive approach. *Trust and Deception in Virtual Societies. Cristano Castelfranchi AND Yao-Hua Tan, Eds., Kluwer Academic Publishers.*

Deutsch, M. (1962). Cooperation and trust. some theoretical notes. *Jones, M.R. ED. Nebraska Symposium on Motivation. Nebraska University Press.*

Golbeck, J. (2005). *Computing and Applying Trust in Web-based Social Networks.* Ph.D. thesis, University of Maryland, College Park, MD, USA.

Golbeck, J. (2006). Generating Predictive Movie Recommendations from Trust in Social Networks.. *Proceedings of The Fourth International Conference on Trust Management.*

Golbeck, J., & Hendler, J. (2006). Filmtrust: Movie recommendations using trust in web-based social networks. *IEEE Consumer Communications and Networking Conference.*

Golebmiewski, R., & McConike, M. (1975). The centrality of interpersonal trust in group processes. *Theories of Group Processes. Cary Cooper Ed. Hoboken, NJ: Wiley.*

Kamvar, S. D., Schlosser, M. T., & Garcia-Molina, H. (2004). The eigentrust algorithm for reputation management in p2p networks. *Proceedings of the 12th International World Wide Web Conference.*

Katz, Y., & Golbeck, J. (2006). Social network-based trust in prioritized default logic. *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 06).*

Levin, R., & Aiken, A. (1998). Attack resistant trust metrics for public key certification.. *7th USENIX Security Symposium.*

Marsh, S. (1994). *Formalising Trust as a Computational Concept.* Ph.D. thesis, University of Stirling, Department of Mathematics and Computer Science.

Massa, P., & Bhattacharjee, B. (2004). Using trust in recommender systems: an experimental analysis. In Jensen, C., Poslad, S., & Dimitrakos, T. (Eds.), *Proceedings of the 2nd International Conference on Trust Management*, Vol. 2995 of *LNCS*, Oxford, UK. Springer-Verlag.

Newcomb, T. (1961). *The Acquaintance Process.* Holt, Rinehart, and Winston, New York, NY, USA.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. *Technical Report 1998, Stanford University.*

Richardson, M., Agrawal, R., & Domingos, P. (2003). Trust management for the semantic web. *Proceedings of the Second International Semantic Web Conference*.

Sinha, R., & Swearingen, K. (2001). Comparing recommendations made by online systems and friends. *DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries*.

Swearingen, K., & Sinha, R. (2001). Beyond algorithms: An hci perspective on recommender systems. *Proceedings of the ACM SIGIR 2001 Workshop on Recommender Systems*.

Sztompka, P. (1999). *Trust: A Sociological Theory*. Cambridge University Press.

Ziegler, C.-N. (2005). *Towards Decentralized Recommender Systems*. Ph.D. thesis, Albert-Ludwigs-Universität Freiburg, Freiburg i.Br., Germany.

Ziegler, C.-N., & Golbeck, J. (2006). Investigating Correlations of Trust and Interest Similarity.. *Decision Support Services*.

Ziegler, C.-N., & Lausen, G. (2004). Spreading activation models for trust propagation. In *Proceedings of the IEEE International Conference on e-Technology, e-Commerce, and e-Service*, Taipei, Taiwan. IEEE Computer Society Press.