On Measuring Social Friend Interest Similarities in Recommender Systems

Hao Ma
Microsoft Research
Redmond, WA 98052
haoma@microsoft.com

ABSTRACT

Social recommender system has become an emerging research topic due to the prevalence of online social networking services during the past few years. In this paper, aiming at providing fundamental support to the research of social recommendation problem, we conduct an in-depth analysis on the correlations between social friend relations and user interest similarities. When evaluating interest similarities without distinguishing different friends a user has, we surprisingly observe that social friend relations generally cannot represent user interest similarities. A user's average similarity on all his/her friends is even correlated with the average similarity on some other randomly selected users. However, when measuring interest similarities using a finer granularity, we find that the similarities between a user and his/her friends are actually controlled by the network structure in the friend network. Factors that affect the interest similarities include subgraph topology, connected components, number of co-friends, etc. We believe our analysis provides substantial impact for social recommendation research and will benefit ongoing research in both recommender systems and other social applications.

Categories and Subject Descriptors: J.4 [Computer Applications] Social and Behavioral Sciences; H.3.3 [Information Search and Retrieval] Information Filtering

Keywords: Friend, Interest Similarity, Recommender Systems, Connected Component, Subgraph Topology

1. INTRODUCTION

Due to its commercial values as well as the research challenges, *Recommender System* has been extensively studied both in industry and academia during the past decade. Recommendation techniques are currently powering many successful online services, including but not limited to product recommendation at Amazon, movie recommendation at Netflix, video recommendation at Hulu, music recommendation at Pandora, etc.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609635 .

Recently, the boom of social networking services fosters the research of trust-aware recommender systems [5, 6, 9, 11, 12]. No matter what techniques are utilized in these approaches, the basic assumption behind these work is that "birds of a feather flock together". In order to evaluate the relationships between trust and user interest similarity, Ziegler et al. in [20] studied hundreds of users observed in two real world datasets, and concluded that their experimental result shows strong indication towards positive interactions between interpersonal trust and user interest similarity in recommender systems.

Although previous preliminary work suggests positive relationship between trust and user interest similarity, many research questions are still left open and need to be further explored.

First of all, previous work on measuring the correlation between trust relation and user interest similarity is performed in a relatively coarse level. In [20], the authors only compared trusted peer similarity with overall peer similarity. Actually, there are many other interesting directions we can further investigate: How does the social peer similarity compare with the random peer similarity? How many social peers will hit one user's Top-N similar user list? How diverse are those social peers in one user's social network?

Secondly, "trust" is only one of many types of social relations. Only few online recommender systems, like Epinions, have the implementation of trust mechanism. On the contrary, many popular recommender systems are designed for online users to interact with their friends in the real life, like Netflix, Flixster, Douban, Foursquare, etc. It is hence important to study the correlations between social friendships and user interest similarities.

Lastly, "trust relationships" are quite different from "social friendships" in many aspects. As mentioned in [10], in a recommender system with "trust" implementation, when a user u likes a review or opinion issued by another user v, user u can add user v to his/her trust list. This process of trust generation is a unilateral action that does not require user v to confirm the relationship. It also indicates that user u does not need to even know user v in the real life. However, "social friendships" refer to the cooperative and mutual relationships that surround us, such as classmates, neighbors, relatives, or colleagues, etc. From the definitions of these two types of social relations, we can see that, in trust-aware recommender systems, one can assume that users may have similar tastes with other users they trust [20]. However, this hypothesis may not be held in friend-based recommender systems since the tastes of one user's friends may vary significantly. Some friends may share similar tastes with this user while other friends may have totally different tastes. Hence, some natural research questions we can explore are: Does friend relationship also indicate positive connection with user interest similarity? If not, can we just simply claim friendship is not related with interest similarity at all?

In this paper, aiming at addressing all the aforementioned research questions, we conduct several in-depth experiments on two large friend communities extracted from real world recommender systems, i.e., Douban friend community and Foursquare friend community. The Douban dataset is composed of a user-item rating matrix as well as the associated friend network, while the Foursquare dataset consists of a user-location check-in matrix and the associated friend network. In order to show the differences between friend community and trust community, in Section 4, we also conduct a comparison analysis for the Epinions trust community. This Epinions dataset also includes a user-item rating matrix as well as the associated trust network.

We observe a range of interesting phenomena occurring in these three communities. The major findings are summarized as follows.

When evaluating interest similarities without distinguishing different social peers a user has, we notice different patterns among these communities.

- We observe strong positive correlation between social trust and user interest similarity. Generally speaking, a user is more similar with his/her trusted peers than randomly sampled users in the community. This conclusion also coincides with Ziegler's previous experiments [20] which are conducted on relatively smaller datasets.

- As to the social friend communities, we find that social friend relationships generally cannot represent user interest similarities in recommender systems. An interesting phenomenon is that the average similarity between a user and his/her friends is even correlated with the average similarity between this user and a set of randomly chosen users in the communities.

- We also notice that, in the social friend communities, a user's similarities with his/her friends are very diverse. This indicates that some friends are quite similar with this user while some other friends are dissimilar with him/her.

In order to detect which friends are more similar with the target user, we measure interest similarities by further exploring the network structures and properties in a user's friend network. Some key observations are:

- The number of co-friends two users share is a factor that can dominate the similarity between these two users. Suppose user u_f is one of the friends user u_i has, then u_i and u_f are more similar if more shared friends are observed between these two users. Quantitatively, in the Foursquare dataset we have, the average similarity between two users who share more than 32 friends is 1.54 times of the average similarity between two users who share no friends.

- The subgraph topology in a user's friend network is another strong indicator that controls the similarity. For the five-node subgraph in one user's Foursquare friend network, the average similarity between this user and five fully connected nodes is 1.60 times of the average similarity between this user and five fully disconnected nodes.

- We also notice that the number of connected components, the size of each component as well as the density in each component can also be used to identify the similarity levels between this users and his/her friends.

We believe that our analysis and findings in this paper provide insightful observations for social recommendation research and will also benefit recommender system designers to develop a more effective platform that can fully utilize the social information.

The remainder of this paper is organized as follows. Section 2 introduces several related work in the literature. Section 3 gives detailed descriptions and statistics of the datasets utilized in this paper. Section 4 shows the comparison analysis with the Epinions trust network, while Section 5 conducts experiments based on the friend network structures and properties. The implications for recommender systems and other social applications are summarized in Section 6, followed by the conclusion and future work in Section 7.

2. RELATED WORK

In this section, we review two research directions which are relevant to our work: user interest analysis in social recommender systems and user interest analysis in other social applications.

2.1 User Interest Analysis in Social Recommender Systems

Taking advantages of the proliferation of online social networking services, the research of social recommender systems becomes more and more popular. Many social-enhanced recommendation algorithms [2, 4, 5, 6, 9, 10, 11, 12, 13, 19] are proposed to improve recommendation quality of traditional approaches. In order to generate better recommendation results, these algorithms either utilize neighborhood-based methods or adopt latent factor models to incorporate social relations. No matter what techniques are developed, the basic assumption employed in these work is that users' social relations can positively reflect users' interest similarities.

Aiming at supporting the assumption mentioned above, in [20], Ziegler et al. conducted preliminary analysis on hundreds of users observed in two trust-based recommender systems. They argued that in order to provide meaningful results, trust must reflect user similarity to some extent because recommendations only make sense when obtained from like-minded people exhibiting similar taste. Their analysis concludes that they found strong indication towards positive correlation between trust and user interest similarity in recommender systems. However, as mentioned in Section 1, there are still many research problems need to be further investigated. In this paper, we provide an in-depth investigation on the study of correlations between social friend relations and user interest similarities in recommender systems.

2.2 User Interest Analysis in Other Social Applications

Although the research of user interest analysis in recommender systems is preliminary and limited by the availability of public datasets, the research of user interest analysis in other social applications [1, 7, 8, 15, 17, 18] is relatively active.

In [8], Leskovec et al. studied 180 million users on Microsoft Messenger social network, and found that people tend to communicate more with each other when they have similar age, language, and location. By jointing the users of Microsoft Messenger with the users of Microsoft search engine, Singla et al. [15] discovered very strong relation between who talks to whom on the instant messaging network, and what they search for. The analysis reveals that people who chat with each other (using instant messaging) are more likely to share interests (their web searches are the same or topically similar). The more time they spend talking, the stronger the relationship is.

In [7], Lee et al. studied the collaborative tagging system CiteULike, and concluded that users connected by social networks exhibit significantly higher similarity on all explored levels (items, meta data and tags) than non-connected users. In [17, 18], Zhen et

Table 1: Statistics of User-Item Rating Matrix of Douban Friend Dataset

Statistics	User	Item
Min. Num. of Ratings	1	1
Max. Num. of Ratings	6,328	49,504
Avg. Num. of Ratings	129.98	287.51

Table 2: Statistics of Social Friend Network of Douban Friend Dataset

Statistics	Friends per User
Min. Num.	1
Max. Num.	986
Avg. Num.	13.07

Table 3: Statistics of User-Location Check-in Matrix of Foursquare Friend Dataset

Statistics	User	Item
Min. Num. of Unique Check-ins	1	1
Max. Num. of Unique Check-ins	324	2,426
Avg. Num. of Unique Check-ins	26.88	10.60

Table 4: Statistics of Social Friend Network of Foursquare Friend Dataset

Statistics	Friends per User
Min. Num.	1
Max. Num.	866
Avg. Num.	12.80

al. presented studies on the quality of inferring user interests from friends in one of the largest global organizations. They demonstrated that there exists large variance of the inference quality when user contributed content considerably varies and the content types are diverse. To allow social applications make informed decisions on when to utilize inferred user interests, they also further investigated relevant factors and presented a method to predict inference quality based on various network features.

In this paper, different from the work in other social applications, we focus on the recommender system domain, and evaluate the correlations between social friend relations and user interest similarities by exploring various factors that can potentially affect the correlations.

3. DATASET DESCRIPTION

We analyze data from online social applications where users not only have social network information, but also have the user preference data, like user-item rating data or user-location check-in data.

3.1 Definitions of Social Relations

Having a good understanding of the social relations we study in this paper will help us better interpret the experimental results. Hence, before we describe the datasets, we first briefly introduce the differences of social relations studied in this paper.

In this paper, we mainly study the social friend relationships given the motivations mentioned in Section 1. The social friend relationships on the web are very close to the real world friendships. Typically, web sites with friendship-building implementations will initially ask users to add friends through their email accounts, and their friends will be asked to confirm the relationships. Hence, a user's online friends will most probably have a large overlap with this user's offline friends. Also, we can see that friend relationships are mutual relationships.

Table 5: Statistics of User-Item Rating Matrix of Epinions Trust Dataset

Statistics	User	Item
Min. Num. of Ratings	1	1
Max. Num. of Ratings	1960	7082
Avg. Num. of Ratings	12.21	7.56

Table 6: Statistics of Social Trust Network of Epinions Trust Dataset

Statistics	Trust per User	Be Trusted per User
Min. Num.	1	0
Max. Num.	1763	2443
Avg. Num.	9.91	9.91

In order to show how different the social friend networks are comparing with the social trust networks, we also include the analysis of a social trust network in Section 4 as the baseline network. Typically, on a web site with trust mechanism, user u will add user v into his/her trust list if user u finds user v has similar taste with him/her through user v's ratings, public comments, reviews, etc., or user u agrees with most of opinions issued by user v. This relationship is unilateral, which means user u trusts user v does not necessarily indicate that user v will also trust user u.

3.2 Douban Friend Dataset

The first data source we choose is Douban[1] dataset. Douban, launched on March 6, 2005, is a Chinese Web 2.0 web site providing user rating, review and recommendation services for movies, books and music. It is also the largest online book, movie and music database and one of the largest online communities in China. Users can assign 5-scale integral ratings (from 1 to 5) to movies, books and music. It also provides Facebook-like social networking services, which allows users to find their friends through their email accounts[2]. This means that most of the friends on Douban actually know each other offline. Hence, Douban is an ideal source for our research on measuring the correlations between social friend and user interest similarity.

Users on Douban can join different interesting groups. At the time when were crawling Douban web site (November 2009), there were more than 700 groups under the "Movie" subcategory. We crawled all the users in these groups, and used these users as seeds to further crawl their social networks with their movie ratings. Finally, we obtain 129,490 unique users and 58,541 unique movies with 16,830,839 movie ratings. As to the social friend network, the total number of friend links between users is 1,692,952. The statistics of the Douban user-item rating matrix and social friend network are summarized in Table 1 and Table 2, respectively.

3.3 Foursquare Friend Dataset

The second dataset we use in this paper is Foursquare[3] dataset. Foursquare is a location-based social networking service for mobile devices. Users can check-in at venues using Foursquare mobile application. Users can also add/invite friends by using email accounts

[1] http://www.douban.com

[2] At the time when we were crawling the Douban dataset, Douban only allowed Facebook-like relationship building approach. Now Douban also supports Twitter-like following mechanism.

[3] https://foursquare.com/

or mobile phone numbers. The friendship building process needs users' mutual agreements. Hence, the data from Foursquare is another source for our research purpose.

This Foursquare dataset we obtain contains 16,748 users who checked in totally 42,460 unique locations. Notice that in this dataset, we do not have rating data since check-in behavior is a binary action. However, many users will check-in the sample location multiple times, which also indicates how much this user likes a location. Thus, in the aggregated user-location check-in matrix, each entry is an integer number that represents a user's check-in frequency on a location. The total number of entries in this user-location check-in matrix is 450,114. As to the social friend network, there are a total of 231,148 friendships observed in this network. Other statistics of the user-location check-in matrix and the user social friend network are summarized in Table 3 and Table 4, respectively.

3.4 Epinions Trust Dataset

The third dataset we utilize is the Epinions[4] trust dataset. Epinions.com is a well known knowledge sharing site and review site that was established in 1999. Online users need to register and begin submitting their own personal opinions on topics such as products, companies, movies, or reviews issued by other users. Users can also assign products or reviews integral ratings from 1 to 5 (5 indicates "like" while 1 indicates "dislike"). These ratings and reviews will influence future customers when they are deciding whether a product is worth buying or a movie is worth watching. Every member of Epinions maintains a "trust" list which presents a network of trust relationships between users. This network is called the "web of trust", and is used by Epinions to re-order the product reviews such that a user first sees reviews by users that they trust. Epinions is thus an ideal source for our analysis on evaluating the relation between trust and user interest similarity.

A user's trust list as well as this user's rating information are publicly available to all the online users. Hence it is very convenient for us to analyze the data on Epinions. The dataset used in our experiments is collected by crawling the Epinions.com site on January 2009. It consists of 51,670 users who have rated a total of 83,509 different items. The total number of ratings is 631,064. Other statistics of the Epinions user-item rating matrix is summarized in Table 5. As to the user social trust network, the total number of issued trust statements is 511,799. The statistics of this data source is summarized in Table 6.

4. COMPARISON ANALYSIS WITH THE TRUST NETWORK

In this section, we give detailed analysis on evaluating the correlations between social relations and user interest similarities without distinguishing different social peers a user has.

4.1 Definition of Similarity

Since every user's interest can be represented by the ratings/check-ins this user has, there are several similarity calculation functions we can borrow in the literature.

In this section, we utilize the Pearson Correlation Coefficient (PCC) [14] as the metric to evaluate the similarity between user i and user j, which is defined as:

$$s_{ij} = \frac{\sum_{p \in I(i) \cap I(j)} (r_{ip} - \overline{r}_i) \cdot (r_{jp} - \overline{r}_j)}{\sqrt{\sum_{p \in I(i) \cap I(j)} (r_{ip} - \overline{r}_i)^2} \cdot \sqrt{\sum_{p \in I(i) \cap I(j)} (r_{jp} - \overline{r}_j)^2}}, \quad (1)$$

where $I(i)$ represents a list of items/locations that user i rated/visited, p belongs to the subset of items or locations which user i and user j both rated or visited, r_{ip} is the rating user i gave to item p or the number of times that user i checked in at location p, and $\overline{r}_i$ represents the average score of user i.

From the above similarity definition, we can see that s_{ij} is ranging from -1 to 1, and a larger value means users i and j are more similar. We employ a mapping function $f(x) = (x+1)/2$ to bound the range of PCC similarities into $[0, 1]$.

We also test a number of other standard similarity measures, including the Vector Space Similarity (VSS) [3] and others. For all of them, we observe similar trends in the analysis, and the results are not qualitatively different. Hence, we only report the results using PCC similarity function.

In the following subsections, we will perform detailed analysis on three different datasets.

4.2 Comparison with Random Users

The first analysis we perform is to understand the research question: how does social peer similarity compare with random peer similarity? More specifically, we conduct the experiments as follows:

1. For each user i, we calculate the average social peer similarity

$$\overline{s}_i = \frac{\sum_{k \in S(i)} s_{ik}}{|S(i)|}, \quad (2)$$

where $S(i)$ represents the list of social peers of user i.

2. We also calculate the average random peer similarity

$$\overline{r}_i = \frac{\sum_{k \in R(i)} s_{ik}}{|R(i)|}, \quad (3)$$

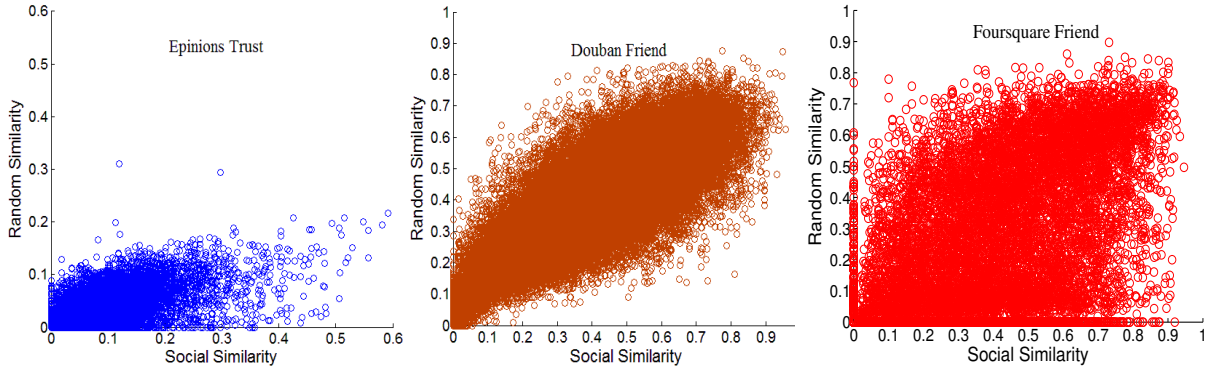
where $R(i)$ represents the list of randomly selected peers for user i, which has the same size with $S(i)$, and $R(i) \cap S(i) = \varnothing$.

3. We then compare the values between social similarity and random similarity for each user in great detail.

The motivation for comparing social similarity with random similarity is that we expect the values of social similarities are much higher than those of random similarities if there is a strong positive correlation between social peers and user interest similarities, and vice versa.

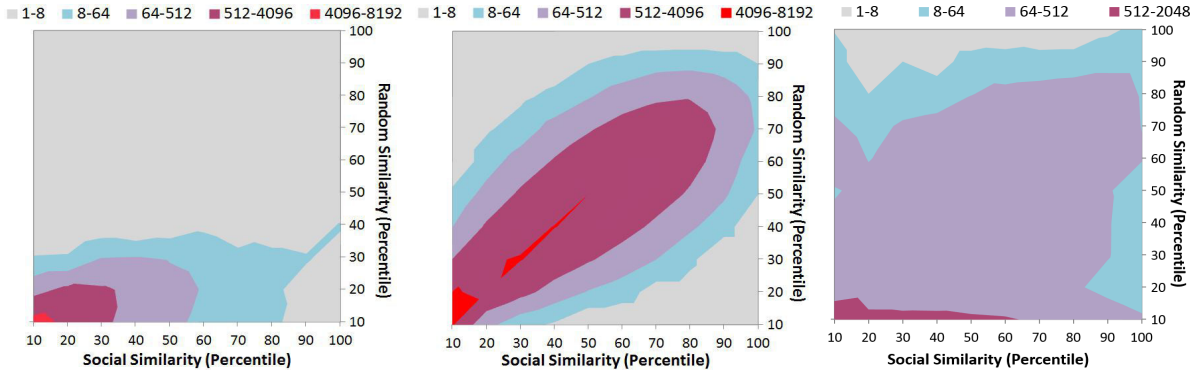
We calculate the social similarity and random similarity for every user in three datasets. In order to reduce noises, we require that each user needs to have at least four claimed social peer relations. For those users whose numbers of social peers are less than four, we do not include them into this analysis. Moreover, we run the random selections several times, and similar patterns are observed. Figure 1 plots the correlations between social similarity and random similarity on three different datasets, respectively. Every data point in the figures represents a user with the x-axis specifying social similarity value and the y-axis indicating the related random similarity value. Figure 2 shows the corresponding heat-map of

[4] http://www.epinions.com



(a) Epinions Trust (b) Douban Friend (c) Foursquare Friend

Figure 1: The Correlations between Social Similarity and Random Similarity



(a) Epinions Trust (b) Douban Friend (c) Foursquare Friend

Figure 2: Heat-maps on the Correlations between Social Similarity and Random Similarity

each sub-figure in Figure 1. The density of colors illustrates the intensity of users. From these two figures, we have the following observations:

1. First of all, in the Epinions trust community, we notice that the plots in Figure 1(a) and Figure 2(a) exhibit strong biases towards the lower-right region, which show a strong indication that social trust information has high correlation with user interest similarity. We will quantify the correlation between social similarity and random similarity later in this section.

2. Secondly, in the Douban friend community, we obtain totally different trends. From Figure 1(b) and Figure 2(b), we actually cannot find evidences that social friend information is correlated with user interest similarity. We notice that the social similarity is even highly correlated with random similarity, which indicates that: in terms of user interest similarity, a user's friends are almost equivalent with a list of other users randomly drawn from the user space. If we connect this conclusion with the formation process of social friend network we described in Section 3.1, we will find this conclusion is actually very reasonable and representative. As mentioned earlier, this Douban friend social network is very close to the real world social friend network. Imagining the real world scenario, actually only very few of your friends have similar tastes with you. This problem is even more severe in online social friend network. For example, on Facebook, a typical user has hundreds of friends, but only those friends who highly interact with this user will probably share similar tastes with this user.

3. Thirdly, in the Foursquare friend community, similar with the Douban friend community, we also cannot find obvious evidences that social friend information is related to user interest similarity in 1(c) and Figure 2(c).

In order to quantify the correlations between social relations and user interest similarities, we would like to measure the proportion of users whose social similarities are greater than their random similarities in each community (i.e., $\overline{s}_i - \overline{r}_i > 0$). We see huge differences between two social friend communities and the social trust community. Totally, in Epinions trust dataset, there are 82.9% of users whose social similarities are greater than their random similarities. However, this number drops to 45.1% and 52.8% in Douban friend and Foursquare friend datasets, respectively. From these numbers, we again observe strong correlations between social trust and interest similarity, while we cannot draw any conclusions between the social friend and interest similarity.

4.3 Top-N Analysis

The second analysis we perform is to see how many a user's social peers will hit this user's Top-N similar neighbors in those three datasets. In an ideal case, if we find most of a user's Top-N similar neighbors come from this user's social network, then we can draw the conclusion that users' social relations are highly correlated with users' interest similarities.

We define user i's Precision at N ($P_i@N$) as follows:

$$P_i@N = \frac{T_N(i) \cap S(i)}{N}, \tag{4}$$

where $T_N(i)$ represents the list of Top-N most similar users of user i, while $S(i)$ specifies the list of social peers of user i. Then the

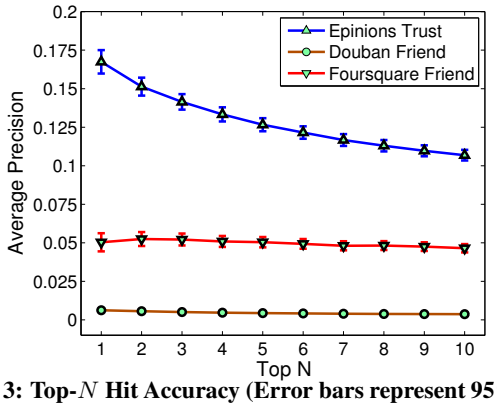


Figure 3: Top-N Hit Accuracy (Error bars represent 95% confidence intervals)

Average Precision at N ($AP@N$) for each dataset can be defined as:

$$AP@N = \frac{\sum_{i=1}^{m} P_i@N}{m}, \tag{5}$$

where m is the number of users in each dataset.

The $AP@N$ analysis with error bars on three datasets is shown in Figure 3. From the results, we can see that the social trust community has the highest $AP@N$ where the scores are much higher than the two social friend communities. This again indicates that social trust information is much more correlated with user interest similarity than the social friend information.

In this Top-N analysis, one may argue that the experiments conducted are unfair for these three datasets since the number of users and the average number of social peers in these three communities are different. Actually, we conduct analysis to normalize the precision by taking into consideration of average number of social peers as well as user sizes of different communities. Eventually, we find the curves are very similar with those presented in Figure 3, thus we do not present the details here.

4.4 Consistence Analysis

The third analysis we are interested in is to address the following questions:

- How consistent are one user's social peers?
- Do the similarities between a user and his/her social peers vary a lot?
- Are there any different patterns among these three communities?

In order to answer the above questions, we evaluate the consistences based on the following two metrics, i.e., Mean Average Distance (MAD) and Root Mean Square Distance (RMSD). The definitions of MAD and RMSD for user i are:

$$MAD = \frac{\sum_{k \in S(i)} |s_{ik} - \overline{s}_i|}{|S(i)|}, \tag{6}$$

and

$$RMSD = \sqrt{\frac{\sum_{k \in S(i)} (s_{ik} - \overline{s}_i)^2}{|S(i)|}}, \tag{7}$$

where s_{ik} is the similarity between user i and user k defined in Equation 1, $\overline{s}_i$ is the average social similarity of user i defined in Equation 2, while $S(i)$ represents the list of social peers of user i.

From the definitions, we can see that we are actually measuring in what extent a user's social similarity s_{ik} will deviate from his/her average social similarity $\overline{s}_i$. If a user's social peer similarities all fall into a small range, then his/her MAD and RMSD will be relatively small, which indicates this user's social peers are very consistent with this user. If we observe a large MAD and RMSD value, then this user's social peers are relatively diverse. Figure 4(a) and Figure 4(b) show the analysis results of MAD and RMSD, respectively. In order to reduce the noises, we only consider those users who have at least four social relations.

We notice that the curves of three datasets illustrate different patterns in both MAD and RMSD figures. The figures reveal that a large portion of users in Epinions trust community has relatively small MAD and RMSD values, which implies that users' social peers are relatively more consistent in Epinions trust community. The MAD and RMSD values in Douban friend and Foursquare friend communities are relatively larger, which presents that users' social peers in these two communities are more diverse. We also notice that quite a few users have very large MAD and RMSD values in the Foursquare friend community. This phenomenon suggests these users' social peers are quite diverse, and relatively speaking, some social peers are very dissimilar with these users but other social peers are very similar with these users. In the next section, we focus on how to detect those users who are very similar with the target user by utilizing rich friend network structure and property information.

5. ANALYSIS BASED ON NETWORK STRUCTURES AND PROPERTIES

In Section 4, all the experiments we conduct indicate that there are no clear correlations between friend relations and interest similarities. However, we also conclude that a user's friends' tastes are very diverse in the sense that some friends may be quite similar with the target user while some other friends are diametrically opposed.

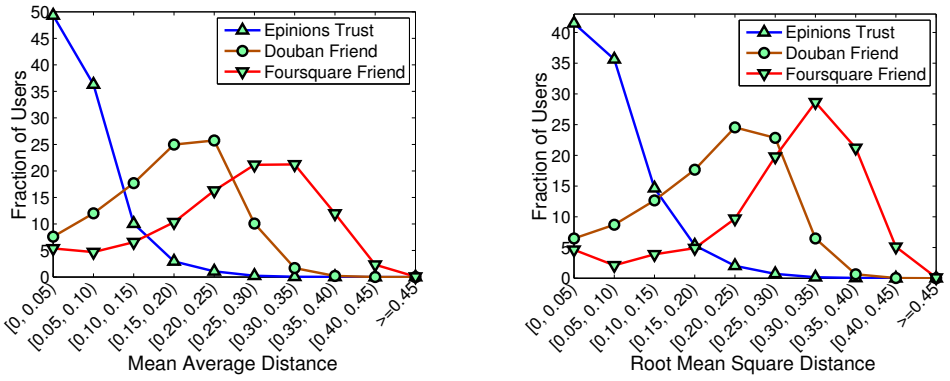
In this section, in order to detect which friends share similar tastes with the target user, we perform several in-depth experiments with finer granularity at different scales by utilizing rich network structure and property information obtained from the friend networks, including the co-friend analysis presented in Section 5.1, the subgraph topology experiments detailed in Section 5.2 as well as the connected components analysis illustrated in Section 5.3.

We exclude the social trust network from this section since we already confirmed that trust is positive correlated with interest similarity and the major focus of this paper is to study the correlation between friend information and user interest similarity.

5.1 Number of Co-friends

This first experiment we conduct in this section is to evaluate how the number of co-friends between two friends can affect the interest similarity between these two friends.

More specifically, for any friend pair in a dataset, we first calculate the similarity between this pair using the similarity function mentioned in Equation 1. Then we count the number of shared friends between this pair. The aggregated results for Douban and Foursquare datasets are shown in Figure 5. In the x-axis, we group the number of co-friends into 8 categories, where "(4, 8]" indicates the number of co-friends is greater than 4 but less or equal to 8. The error bars in this figure represent 95% confidence intervals. Moreover, for all the other figures we present in the following subsections, the error bars are all 95% confidence intervals. Also notice that some errors may be very small, hence the corresponding error bars are barely visible.



(a) Mean Average Distance (MAD) (b) Root Mean Square Distance (RMSD)

Figure 4: Consistence Analysis

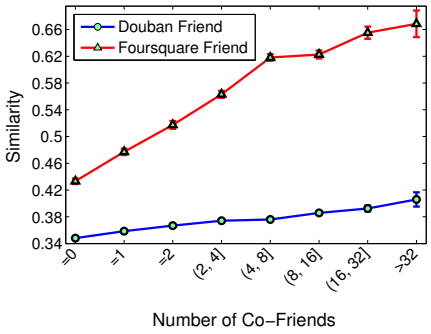


Figure 5: Similarities Conditioned on the Number of Co-Friends

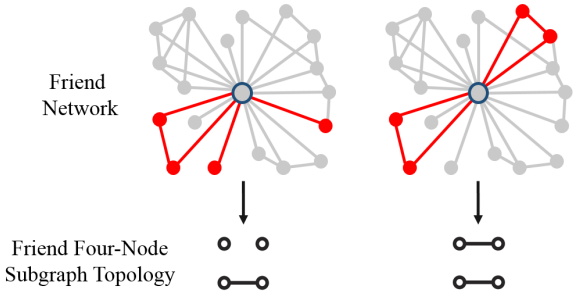


Figure 6: Examples on Friend Subgraph Topology

From Figure 5, we can see both datasets show that two friends are more similar if there are more shared friends between them. In the Foursquare data, the average similarity between two users who share more than 32 friends is even 1.54 times of the average similarity between two users who share no friends. One interpretation for this observation is that, in the real life, if two users share a large amount of friends, then most probably, these two users have similar ages, attended the same highschool/college together, or working on similar fields/topics, etc. All these background information already hints that these two users may have similar tastes.

Moreover, one may observe that interest similarity between a user and his/her friends increases sharply with the number of co-friends for Foursquare dataset, whereas the variation is less pronounced in Douban dataset. This phenomenon is actually very reasonable since as shown in Section 3, different datasets have distinct statistics. There are many factors can affect the ranges of the similarities in two different datasets, including average number of items co-rated, rating patterns, etc. Nevertheless, we can still draw the conclusion that the number of co-friends is a strong signal that controls user interest similarity.

5.2 Subgraph Topology

The co-friend analysis in Section 5.1 measures the similarity between a user and one of his/her friends each time. In this section, motivated by the work in studying the social contagion problem [16], we are also interested in evaluating the similarity between a user and a subset of his/her friends.

Figure 6 shows two examples on how we construct the four-node friend subgraphs. In each example, the middle node represents the target user, while all the other nodes around this user are the friends of this user. The highlighted four nodes as well the edges between these four nodes form a subgraph. Figure 6 illustrates two different subgraph topology patterns: the left subgraph has one edge and three components while the right one has two edges and two components.

The fundamental question we investigate here is the following: how does the average similarity between the target user and the subgraph nodes depend on the different subgraph topology patterns from his/her friend network?

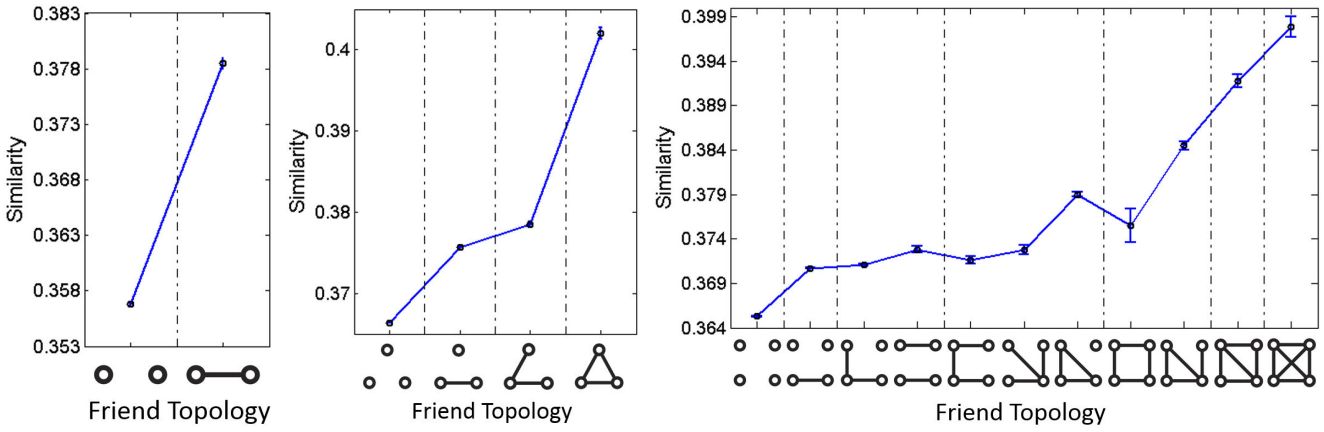
In this paper, we conduct analysis on two-node, three-node, four-node and five-node subgraphs, respectively. There are totally 2 distinct topology patterns in two-node subgraphs, 4 in three-node subgraphs, 11 in four-node subgraphs and 32 in five-node subgraphs.

When constructing all the subgraph patterns for a user, it is infeasible to enumerate all the possible subgraphs due to the following reasons: (1) When a user has many friends, enumerating all the possible n-node subgraphs is very time consuming. (2) If a user has a large number of friends, the generated huge number of subgraphs for this single user can possibly dominate the overall distribution, which will result in an unfair analysis. Hence, without loss of generality, for each user, we randomly sample a certain amount of subgraphs. In this paper, we set the sample number for each user to 50,000.

The trends of this analysis are presented in Figure 7, Figure 8 and Figure 9, respectively.

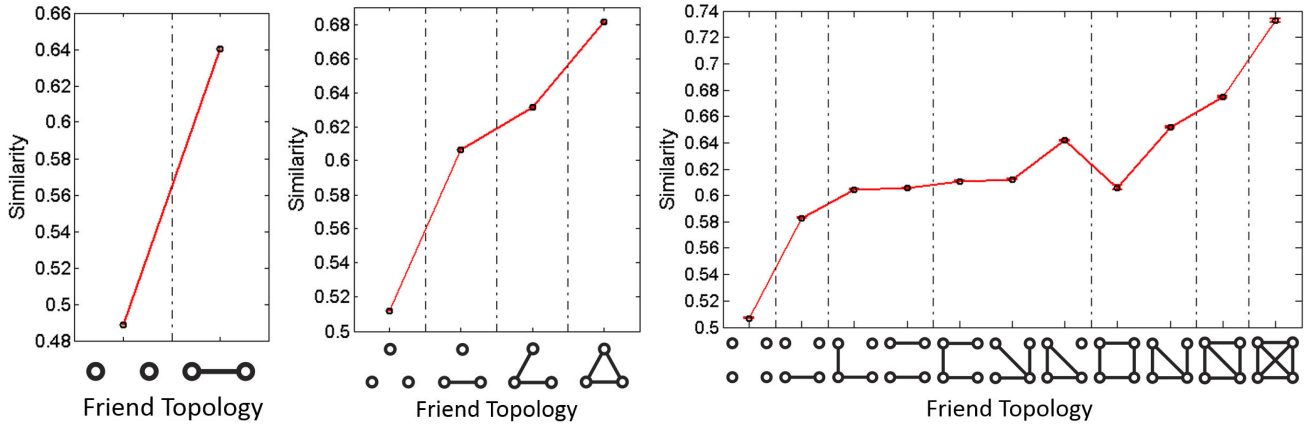
Figure 7 and Figure 8 summarize results for the two-node, three-node and four-node topology patterns in Douban and Foursquare datasets. From these two figures, we can see that the similarity is largely controlled by the number of edges in each topology patterns. For example, in the four-node subgraph, when more edges in a certain topology pattern are observed, a user's average similarity with those four nodes is more similar.

We also examine the similarity analysis results on the Foursquare dataset conditioned on all the five-node topology patterns in Figure 9. The results on Douban dataset also share similar trends with Foursqure dataset, hence we do not present details in this paper due to the space limitation.



(a) Two Nodes (b) Three Nodes (c) Four Nodes

Figure 7: Similarities Conditioned on the Friend Topology (Douban Friend)



(a) Two Nodes (b) Three Nodes (c) Four Nodes

Figure 8: Similarities Conditioned on the Friend Topology (Foursquare Friend)

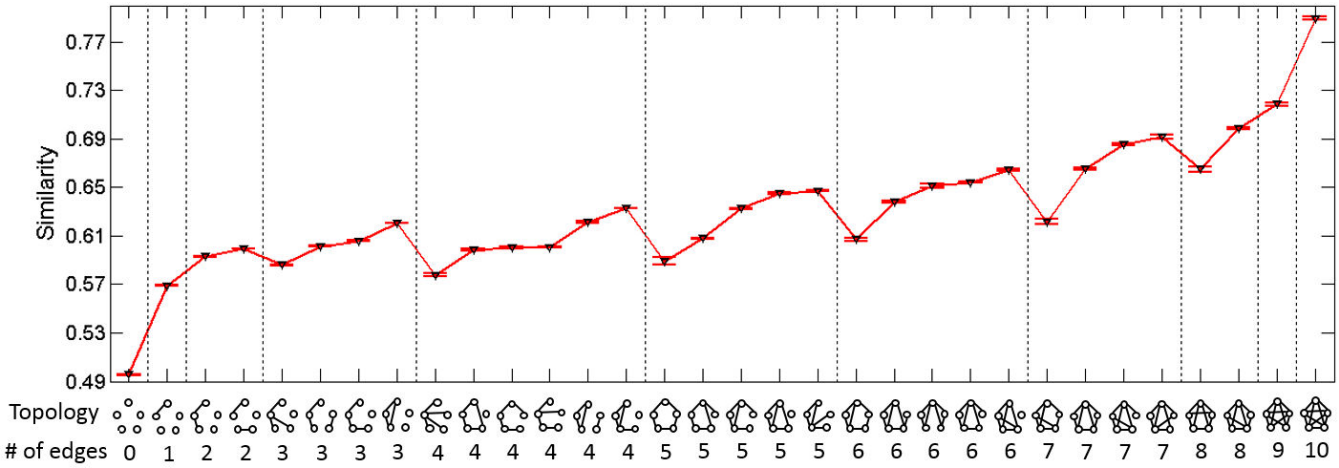


Figure 9: Similarities Conditioned on the Friend Topology (Foursquare Friend, Five Nodes)

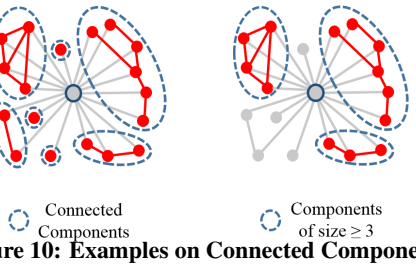


Figure 10: Examples on Connected Components

In Figure 9, we group each topology patterns by the number of edges in each pattern. The vertical dashed lines separate different groups for clearer comparison.

From this figure, we have the following key observations.

First of all, the trend we observe from Figure 7 and Figure 8 still holds. Generally speaking, when more edges are presented, the similarity values are higher. The average similarity between a user and five fully connected friends is 1.60 times of the average similarity between a user and five fully disconnected friends.

Secondly, we also notice that even in the same group with the same edge number, some patterns are larger than other patterns in terms of similarities. Moreover, not every patterns in one group with n edges have larger similarities than some patterns in another group with less edges. It seems that there are some other factors could affect user interest similarities besides the number of edges in each topology pattern.

Thirdly, we see an interesting phenomenon in each group with the same edge number. That is, in each group, the patterns with more connected components generally have larger similarities than those patterns with less connected components. This observation points out that connected components could be another predictor that affects user interest similarities.

5.3 Connected Components

In this section, as motivated by Section 5.2, we study the impact of connected components on measuring the social friend interest similarities.

The left part of Figure 10 demonstrates the concept of connected components, while the right part of this figure shows the connected components of size greater than or equal to 3.

The first experiment we conduct is to evaluate how the number of connected components in a user's friend network can change the average similarity between this user and all his/her friends, as illustrated in Figure 11. We group the number of connected components into 9 categories. If a user's friend network has 31 connected components, then the average similarity between this user and all his/her friends will be grouped into the category "(16, 32]". From this figure, we can see that simply counting the number of connected components leads to a muddled view of predicting user interest similarity. In both datasets, at the beginning, the number of connected components positively affects user interest similarity. However, when this number passes certain threshold, the number of connected components start to negatively impact user interest similarity.

From the above analysis, we conclude that simply using the number of connected components may not be a very effective predictor. Hence, motivated by the observations we obtain in Section 5.2 and this section, we address connected components with finer granularity. That is, we consider the edge density in each connected component.

For each connected component in a user's friend network, we first calculate the average similarity between this user and all the nodes in this connected component. Then we correlate this average

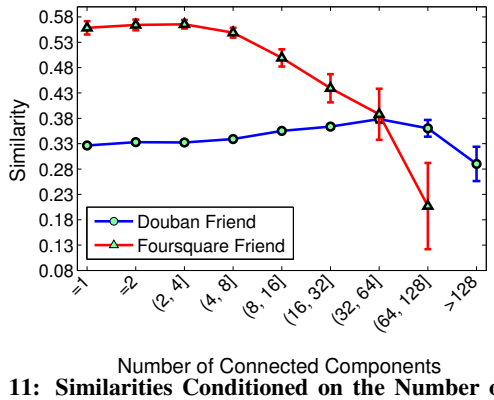


Figure 11: Similarities Conditioned on the Number of Connected Components

similarity with the edge density in this connected component. The aggregated results for Douban and Foursquare datasets are shown in Figure 12(a) and Figure 12(b), respectively.

From these two figures, if we consider simply the number of components of size k or larger, we see that small values of k (like size 3) are not enough; but when k is increased to make the selection over components sufficiently astringent (for example, when we count only components of size 5 or larger), we can see that this metric becomes a significant positive predictor for evaluating user interest similarities.

6. IMPLICATIONS

From Section 4 and Section 5, we obtain many interesting observations. The implication and knowledge we learn from these observations can be exploited by diverse applications which rely on user interest modelings.

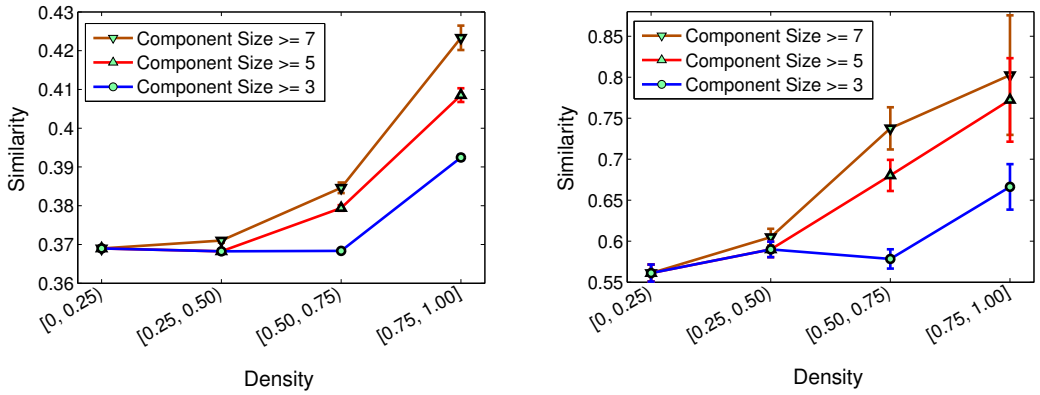
6.1 Implications for Recommender Systems

In recommender systems, we confirm that trust information is a very ideal source to represent users' interests from our analysis based on explicit rating similarity. Moreover, our analysis provides strong support to those trust-aware collaborative filtering methods [6, 9], and interprets why utilizing social trust information can increase the recommendation prediction accuracy.

In terms of social friend information, when treating every friend a user has equally, we cannot find any correlations between social friend and user interest similarity. However, we conclude that using network structures and properties as different contexts, we can find many factors that can positively predict user interest similarity. This observation points that when designing computational social recommendation techniques, all those contexts we analyzed in this paper can be either used as features or treated as motivations to better model the social recommendation problem. These finds could also benefit user experience researchers as well as user interface designers to design a better mechanism in interpreting recommendation results using social contextual information.

6.2 Implications for Other Applications

There are many other applications where social network information plays an important role. The most natural and important application is probably the social search problem. Recently, both Google and Bing released interesting features that are related to social search. The basic idea is to allow contents from users' social network to be surfaced to users' search results or the social side bars. However, based on our investigation, users may not share similar tastes with most of their friends, hence they may not be interested in the social search results which are presented to them. In



(a) Douban Friend (b) Foursquare Friend

Figure 12: Similarities Conditioned on the Graph Densities of Connected Components

some cases, users may find the recommended social results annoying. Thus, in order to improve the user experience, it is very crucial to identify who are the "closest" friends for a given user. Based on all the findings presented in this paper, we can easily help solve this problem by looking at different social contexts we developed.

7. CONCLUSION AND FUTURE WORK

In this paper, we investigate the correlations between social friend and user interest similarity in the context of recommender systems. We found several interesting phenomena that present different patterns and trends among social friend-based recommender systems. We believe our findings provide substantial impact for social recommendation research and will benefit ongoing research in both recommender systems and other social applications.

We still have plenty of tasks we can perform in the future. In this paper, we only evaluate user similarities between directly connected users. We can also further analyze the similarities between users who are multi-hop connected. We believe this will give us another point of view on understanding the correlations between social relationships and user interest similarities.

There is another very important type of social network we do not mention in this study, i.e., social following network. We believe this kind of social networks will have other distinct characteristics. We plan to conduct this analysis in the future.

8. REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Effects of user similarity in social media. In *Proc. of WSDM '12*, Seattle, Washington, USA, 2012.

[2] P. Bedi, H. Kaur, and S. Marwaha. Trust based recommender system for the semantic web. In *Proc. of IJCAI '07*, pages 2677–2682, Hyderabad, India, 2007.

[3] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. of UAI '98*, 1998.

[4] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *Proc. of ICWSM '12*, 2012.

[5] M. Jamali and M. Ester. Trustwalker: a random walk model for combining trust-based and item-based recommendation. In *Proc. of KDD '09*, pages 397–406, Paris, France, 2009.

[6] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proc. of RecSys '10*, pages 135–142, Barcelona, Spain, 2010.

[7] D. H. Lee and P. Brusilovsky. Social networks and interest similarity: the case of citeulike. In *Proc. of HT '10*, pages 151–156, Toronto, Ontario, Canada, 2010.

[8] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proc. of WWW '08*, pages 915–924, Beijing, China, 2008.

[9] H. Ma, I. King, and M. R. Lyu. Learning to recommend with social trust ensemble. In *Proc. of SIGIR '09*, pages 203–210, Boston, MA, USA, 2009.

[10] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *Proc. of WSDM '11*, pages 287–296, Hong Kong, China, 2011.

[11] P. Massa and P. Avesani. Trust-aware collaborative filtering for recommender systems. In *Proc. of CoopIS/DOA/ODBASE '04*, pages 492–508, 2004.

[12] P. Massa and P. Avesani. Trust-aware recommender systems. In *Proc. of RecSys '07*, pages 17–24, Minneapolis, MN, USA, 2007.

[13] J. O'Donovan and B. Smyth. Trust in recommender systems. In *Proc. of IUI '05*, pages 167–174, San Diego, California, USA, 2005.

[14] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proc. of CSCW '94*, pages 175–186, Chapel Hill, North Carolina, United States, 1994.

[15] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *Proc. of WWW '08*, pages 655–664, Beijing, China, 2008.

[16] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proc. of National Academy of Sciences*, 109:5962–5966, April 2012.

[17] Z. Wen and C.-Y. Lin. On the quality of inferring interests from social neighbors. In *Proc. of KDD '10*, pages 373–382, Washington, DC, USA, 2010.

[18] Z. Wen and C.-Y. Lin. Improving user interest inference from social neighbors. In *Proc. of CIKM '11*, pages 1001–1006, Glasgow, Scotland, UK, 2011.

[19] Q. Yuan, L. Chen, and S. Zhao. Factorization vs. regularization: fusing heterogeneous social relationships in top-n recommendation. In *Proc. of RecSys '11*, pages 245–252, Chicago, Illinois, USA, 2011.

[20] C.-N. Ziegler and J. Golbeck. Investigating correlations of trust and interest similarity. *Decis. Support Syst.*, 43:460–475, March 2007.