On the Quality of Inferring Interests From Social Neighbors

Zhen Wen
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532 USA
zhenwen@us.ibm.com

Ching-Yung Lin
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532 USA
chingyung@us.ibm.com

ABSTRACT

This paper intends to provide some insights of a scientific problem: how likely one's interests can be inferred from his/her social connections – friends, friends' friends, 3-degree friends, etc? Is "Birds of a Feather Flocks Together" a norm? We do not consider the friending activity on online social networking sites. Instead, we conduct this study by implementing a privacy-preserving large distribute social sensor system in a large global IT company to capture the multifaceted activities of 30,000+ people, including communications (e.g., emails, instant messaging, etc) and Web 2.0 activities (e.g., social bookmarking, file sharing, blogging, etc). These activities occupy the majority of employees' time in work, and thus, provide a high quality approximation to the real social connections of employees in the workplace context. In addition to such "informal networks", we investigated the "formal networks", such as their hierarchical structure, as well as the demographic profile data such as geography, job role, self-specified interests, etc. Because user ID matching across multiple sources on the Internet is very difficult, and most user activity logs have to be anonymized before they are processed, no prior studies could collect comparable multifaceted activity data of individuals. That makes this study unique. In this paper, we present a technique to predict the inference quality by utilizing (1) network analysis and network autocorrelation modeling of informal and formal networks, and (2) regression models to predict user interest inference quality from network characteristics. We verify our findings with experiments on both implicit user interests indicated by the content of communications or Web 2.0 activities, and explicit user interests specified in user profiles. We demonstrate that the inference quality prediction increases the inference quality of implicit interests by 42.8%, and inference quality of explicit interests by up to 101%.

KDD'10, July 25–28, 2010, Washington, DC, USA.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

User interest modeling, Social networks, Quality

1. INTRODUCTION

Modeling user interests is important for search and recommender systems to provide personalized results to meet individual user needs [25]. Towards this goal, existing works have studied a user's explicit interests specified in his profile, or implicit interests indicated by his prior interactions with various types of information, such as content the user has created or read including web pages, documents and email. Recently, the proliferation of online social networks spark an interests of leveraging social network to infer user interests [26], based on the existence of social influence and correlation among neighbors in social networks [21]. For applications that can directly observe a user's behavior (e.g., logs of search engines he uses), inferring interests from his friends in social networks provides one extra useful enhancement. For many other applications, however, it is difficult to observe sufficient behavior of a large number of users. In such scenarios, inferring their interests from their friends can be the only viable solution. For example, for a new user in a social application, the application may only have information about his friends who are already using it. To motivate the new user to actively participate, the application may want to provide personalized recommendations of relevant content. To this end, the application has to infer his interests from friends.

However, there exists huge variation in the types and amount of information in social interactions. According to existing studies on enterprise social networks [5], only a small percentage of employees (e.g., $< 10\%$) may actively contribute social content using one or more social software (e.g., blogs, social bookmarking and file sharing). But a large number of other employees (e.g., $> 90\%$) may seldom do so. Moreover, certain user contributed data may not be accessible (e.g., private files) or can not be associated with a particular user (e.g., anonymous data). That results in both a demand and a challenge for accurate user interest modeling, especially for *inactive users* in social networks,

i.e., users that do not contribute social content. On one hand, accurate user interest modeling can provide personalized search and recommendation results, and thus may help to increase the usage of social software. On the other hand, the available observations of users are sparse and exist in multiple types of media.

There resides more information about social relationships via traditional communication media, such as emails, instant messaging, meeting calendars, etc. Comparing to social software, they provide better accuracy in inferring social networks, since most people still spend significant time of these media. However, these data are more private, and thus, seldom could prior researches in the literature utilize such information without potential infringing privacy or collaboration from the communication service provider, which is actually illegal in some countries. Fortunately, we solved the privacy issue by designing a rigorous distributed social sensing system [16] and went through lawyer, privacy officer, and union reviews in each country to make sure the system follows the privacy and data protection laws. Up to Jan 2010, this system has been approved worldwide except 5 European countries.

We have been capturing the multifaceted activities of 30K+ people, including communications such as emails, instant messaging, meetings, etc, and web 2.0 activities, such as blogs, wiki, file sharing, and social bookmarking for up to 3 years in more than 70 countries. After anonymizing the identity and the content of these data, we are able to quantitatively infer the social networks of 400K employees within the organization. Figure 1 illustrates the three top social media[1] with the largest contributor populations, and the relationships among their contributor populations. Overall, the three top social media cover 31K employees out of the 400K employees of the enterprise. With the abundant information of people available, we thus can study the scientific questions of "whether one's interests are related to the people surrounding him?", "how accurately can one's interests be inferred through social networks?", etc.

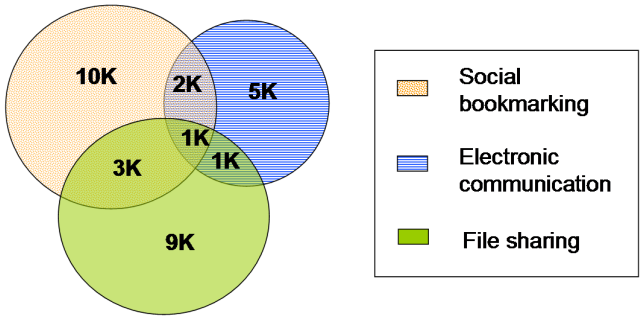

Figure 1: A Venn diagram showing the overlapping contributor populations of three top social media.

To address such challenges, we present a network autocorrelation model based approach that combines multiple types of social media to infer user interests from social neighbors, especially for *inactive users*. In order not to diminish the quality of personalized search and recommendation results by inaccurately inferred user interests, we propose a computational method to measure inference quality (e.g., accuracy of inferred interests) based on observable features such as the user's social network characteristics. Our work offers two unique contributions:

- We demonstrate that combining multiple types of social media significantly improves the quality of inferring users' interests from their friends by as much as 52%, comparing to using only single source.
- We propose a technique to predict the inference quality purely based on network topology characteristics. This allows applications to reliably decide when to infer interests from friends, which can improve the inference quality by up to 101%. This technique is validated on both implicit user interests indicated by user communication content and accessed documents/webpages, and explicit user interests specified by themselves in their profiles.

2. RELATED WORK

Our study on the quality of user interest inference from social neighbors is related to previous approaches to user modeling, and prior work on social influence and correlation among friends in social networks.

2.1 User Modeling

Modeling user interests is common practice for search and recommendation systems. Users may express their interests explicitly in their profiles. Alternatively, user interest models can be derived by analyzing a user's *own* behavior observed from various sources. One type of sources is user authored content such as user generated tags [20, 23] and papers written by users [22]. In addition, user interaction history during web search [1] and indexed desktop documents [25] are used to model user interests. Recently, researchers have started to leverage social cues to enhance user interest modeling. For example, assuming users visited a same page share similar interests, White *et al.* [19], Piwowarski and Zaragoza [26] augment a user's interest model by combining the interests of other users that also visit the same page. Similarly, collaborative filtering (CF) systems model user interests by assuming that those who had similar opinions on a set of items tend to agree again on other items [10]. In these works, they concentrate on one type of user generated data (e.g., web search logs) and the data are available for every user whom they build interest models for. Moreover, they only consider social cues from neighbors with one-degree of separation. In contrast, our study focuses on users in social networks where there are multiple types of data and large variation in user contribution. For example, studies have shown that only a small percentage (e.g., $< 10\%$) of users contributed data in such environments [5]. Our work is the first to study the quality of inferring user interests from friends in such situations. Moreover, in our work we consider friends with various ties and different degrees of separation.

2.2 Social Influence and Correlation

Social influence and correlation among people's behavior have been extensively studied by social network researchers [14]. Lately, the correlation in online social networks is reported by Singla and Richardson [21]. Researchers have explored data mining techniques to model various aspects of social influence. For example, Crandall *et al.* [8] studied the interaction between social influence and *selection*, a pro-

[1]Note that electronic communication can also be considered as a social media, though it is not a Web 2.0 social media.

cess that people tend to form new links to others who are already like them. While social influence is an important factor that induces correlation among people, it is not the only one. Thus, Anagnostopoulos *et al.* [2] present a model to distinguish social influence from other factors of social correlation. Moreover, Tang *et al.* [24] model the dynamics and topic dependency of social influence. While these works provide us with a theoretical foundation to draw upon, our study focuses on the quality of modeling user interests from social neighbors, an important issue for social applications due to practical constraints, such as limited observations with large variations and multiple types.

In other work, several approaches have developed tools for various applications based on social influence and correlation. For example, Liben-Nowell and Kleinberg [15] exploit social correlation to predict future links among users. Ma *et al.* [17] leverage the similarity among ratings from friends to improve items recommendation (e.g., movies, books, etc). Our work similarly makes use of social correlation to infer user interests from social neighbors. However, in contrast to previous work, we focus on understanding the large variance of the inference quality and how the quality can be predicted by social network characteristics.

3. DATASET

We analyze the information content in the electronic communication social networks inside a global information technology firm with more than 400,000 employees. The data contain people's electronic communication data, Web 2.0 social content and employee job-related information.

The people's electronic communication data include email and instant messaging activities. We collected detailed electronic communication records of 8952 volunteer employees in more than 70 countries. To preserve privacy, the original textual content of an email or text message is not saved. Instead, the content is represented as a vector containing the terms appeared in the text as well as their counts after stop-words removal and stemming. In addition, to construct a view of the network that reflects the real communications, we eliminate spam and mass email announcements and are left with 20 million emails and text messaging chats. Because a volunteer's communication may include non-contributors, our system can thus infer the connections and the amount of communications to the non-contributors. From the 8952 volunteers' communication data, we derive the social networks of more than 400K people within the firm, where the strength of a link is computed based the amount of communication between two people (see Equation 5 for the detailed link strength definition).

We also collected social content in various Web 2.0 social software within the enterprise. Among the 400K employees, nearly 16K people created 400K social bookmarks of web pages [18]. Each bookmark includes the web page url and a set of tags. A different but *overlapping* set of 14K people shared 140K files, of which 20K files are public files that we can analyze. Two other social media that have smaller number of users are blogs with 5K contributors and 112K entries, and audio/video media library with 4.2K users and 61K media files. In this study, we focus on social bookmarking and file sharing, the two Web 2.0 social media that have the widest coverage in the enterprise.

Combining the users who contributed to one or more of the three sources, we have 31K employees in total (see Figure 1). After excluding people who only contributed private files, the dateset contains 25315 people. Each person's data include his emails and instant messages, the social bookmarks he has created, and public shared files he has accessed. We denote the sources as *EML*, *BMK*, and *FILE*. In addition, we collected employee job-related information, such as the division an employee belongs to, his job role within the division, and the revenue he generates. Moreover, we collected optional information preferences specified in a subset of employees' profiles, which are used to customize information (e.g., news) displayed on their personal Intranet homepage. To protect the privacy of the employees, their identities are replaced with hash identifiers.

4. USER INTEREST MODEL

In our study, we examine two types of user interests: (1) implicit interests indicated by user content; and (2) explicit interests directly specified by users in their profiles.

4.1 Topic-based Implicit User Interest Model

Because users' contributed content reveal their implicit interests, we model user interests as a set of latent topics extracted from their communication data and contributed social content. We use Latent Dirichlet Allocation (LDA) [4], a generative probabilistic model to extract topics. Given a document corpus D, LDA models each document d as a finite mixture over an underlying set of topics, where each topic t is characterized as a distribution over words. A posterior Dirichlet parameter $\gamma(d,t)$ can be associated with the document d and the topic t to indicate the strength of t in d. As a result, the document d can be reduced to a vector $\vec{\gamma}_d = \langle \gamma(d,t_1), \gamma(d,t_2), ..., \gamma(d,t_T) \rangle$, where T is the total number of topics. Next, we empirically choose $T = 1200$, in order to balance the need to cover the diverse content and the computational complexity.

After topic extraction, we define a $U \times T$ matrix $\mathbf{S}$ to describe user interests using topics, where U is the total number of employees. An element s_{ij} in $\mathbf{S}$ denotes the degree the i-th employee is interested in the j-th extracted topic. We compute s_{ij} by aggregating the strengths of the topic in all of the i-th employee's content. Specifically, we have $s_{ij} = \sum_{d \in D_i} \gamma(d,t_j)$ where D_i is the set of content by the i-th employee and $\gamma(d,t_j)$ is a posterior Dirichlet parameter describing the j-th topic strength in a document d. We then normalize $\mathbf{S}$ by $s_{ij} = \frac{s_{ij}}{\sum_j s_{ij}}$.

Furthermore, since different information sources have different characteristics, we adjust the weight s_{ij} based on its information source by: $s_{ij} = \omega \cdot s_{ij}$, where ω is a constant associated with the information source. For example, social bookmarks are more direct and concise indication of user interests than shared files. Thus, the weight of social bookmarks should be higher than that of file sharing. We empirically set the weights to be $\omega_{BMK} = \omega_{EML} = 0.4$, $\omega_{FILE} = 0.2$.

Finally, practical search and recommendation applications may demand a keyword-based representation of a user's top-N interests. For the top-N interests, the user should be highly interested in them (based on s_{ij}). Moreover, the top-N interests should be distinct to each other (based on word overlap). To incorporates the two constraints, we define an

objective function for top-N interests $\mathbb{T}_N$:

$$obj(\mathbb{T}_N) = \frac{\sigma_1}{N}\sum_{j=1}^{N} s_{ij} - \frac{\sigma_2}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\cos(t_j, t_k)$$

where t_j is the j-th topic, $\cos(t_j, t_k)$ is the cosine similarity, the weights $\sigma_1 = \sigma_2 = 0.5$ but can be adjusted in different application scenarios. We use a greedy method to select the top-N interests. Then, for each interest we output its top N_w keywords (e.g., $N_w = 20$).

4.2 Explicit User Interest Model

User interests can also be explicitly specified in their profiles. In this organization, employees are encouraged to specify information preferences indicating their interests in their online profiles. The interests are specified using terms from a manually designed work-related taxonomy. Overall, 1120 distinct terms are used in explicit interests. We observe that only 29% of employees (116K out of 400K employees) specified their interests, and on average each person specified 10 terms. In addition, many people do not update the interests in years. We use a vector of terms as a user's interest model. The default weight of a term is 1, since the terms specified in profiles are neither weighted nor sorted.

We further examine the relationship between the implicit and explicit user interests. Out of the 25315 users that contributed social content, 8005 of them also explicitly specified interests in their profiles. For these 8005 users that have both interest models, we compare the terms in their explicit interests and the top-20 implicit interests extracted from the contributed content. On average, 60.4% of the terms in a user's explicit interests are covered by the terms in his top-20 implicit interests. In contrast, only 2.2% of the terms in a user's top-20 implicit interests are covered by his explicit interests. This is understandable because the implicit interests are extracted from a large amount of content, while explicit interests are from a limited taxonomy and many users may just specify a small set of of their interests. An in-depth investigation of the relationship between the two types of interests is beyond the scope of this study. Instead, we focus on validating our findings on these two different types of user interest models.

5. INFERRING USER INTERESTS

Users' behavior and interests are influenced by their neighbors in social networks. In this section, we present our approach to inferring user interests using a social influence model based on network autocorrelation. Our approach may especially be useful for inferring interests of inactive users whose behavior is difficult to observe.

5.1 Network Autocorrelation Model

In social network sciences, network autocorrelation models have been widely used to describe the social actors' interdependency as a result of social influence [14]. In these models, people are assumed to establish their own behavior as a result of a diffusion process, in which they appropriately take into account the opinions and behaviors of their significant social neighbors. We adopt this model in our study because only partial observations are available in our network, which makes it difficult to obtain sufficient samples for a large number of inactive users to train a probabilistic graph model such as the one in [24].

Specifically, let $\mathbf{Z}$ be a $(U \times N)$ matrix of values of N endogenous variables for U actors in a network, and let $\mathbf{W}$ denote a $(U \times U)$ matrix where an entry w_{ij} denotes the influence actor j has on actor i. The network autocorrelation model represents ego's attribute as a weighted version of the attributes of his social neighbors:

$$\mathbf{Z} = \rho\mathbf{W} \cdot \mathbf{Z} + \epsilon \tag{1}$$

where ρ is a constant and ϵ is an error term. The entry w_{ij} in the weight matrix $\mathbf{W}$ is defined as a function of the social distance between actor i and actor j. Depending on the characteristics of a given network, multiple types of functions (e.g., exponential function) have been used to construct $\mathbf{W}$. However, the generally appropriate type of function for $\mathbf{W}$ does not exist.

5.2 Inferring Interests from Neighbors

To infer users' interests, we let variable $\mathbf{Z}$ denote the degree users are interested in a set of subjects, where z_{ij} is the i-th user's interest in the j-th subject. The subject can be a topic in the implicit user interests or a term in the explicit user interests. For topic-based implicit user interests, we have $z_{ij} = s_{ij}$, where s_{ij} is an entry of the matrix $\mathbf{S}$ defined in Section 4.1. In our study, we focus on using the normalized s_{ij} value to represent the degree that a user is interested in a subject. Thus we set the network autocorrelation model constant $\rho = 1$. We also ignore ϵ, since the uncertainty of the inference is assessed by the quality of the inferred interests described in Section 5.3 and 5.4. We estimate the degree the i-th user is interested in the j-th subject as

$$z_{ij} = \sum_{k=1}^{U}(w_{ki} \cdot z_{kj}) \tag{2}$$

where z_{kj} is initially set to 0 if the k-th user is inactive so that his interest can not be observed. We further define the weight w_{ki} as an exponential function of the social distance

$$w_{ki} = \exp(-dist(k, j)) \tag{3}$$

where $dist(k, j)$ is the social distance between user k and user i. It is defined by considering the degree of separation in their communication and the amount of communication [16, 27]. Specifically, it is calculated as

$$dist(i, j) = \sum_{k=1}^{K-1}\frac{1}{strength(v_k, v_{k+1})} \tag{4}$$

where $v_1, ..., v_k$ are the nodes on the shorted path from user i to user j, and $strength(v_k, v_{k+1})$ measures the strength of communications between v_k and v_{k+1} and is normalized to vary between 0 and 1. The measure is defined as:

$$strength(i, j) = \frac{log(X'_{ij})}{\max_j log(X'_{ij})} \tag{5}$$

where X'_{ij} is as:

$$X'_{ij} = \begin{cases} 10 & : \texttt{ if } X_{ij} \leq 10 \\ X_{ij} & : \texttt{ otherwise} \end{cases}$$

Here X_{ij} is the total communications between user i and j. This measure of communication strength has been extensively tested and is shown to accurately reflect the strength of tie between users [16].

To infer a user's interests using Equation 2, we only consider the neighbors in his three-degree ego network, i.e., the network of people within three degrees of separation to him. The reason is that previous empirical studies have been able to effectively use the information from neighbors within three degrees of separation for social influence study [7] and social search [16]. Therefore, we focus on the three-degree ego network of a user for inferring his interests.

5.3 Quality of Inferring Implicit Interests

To measure the quality of the inferred implicit interests, we define the quality as: $C = \frac{1}{N}\sum_{j=1}^{N} \max_{t' \in \mathbb{T}'_N}[\cos(t_j, t')]$, where t_j is the j-th topic in the inferred top-N interests, $\mathbb{T}'_N$ is the ground-truth top-N interests, and $cos(t_j, t')$ is the cosine similarity. Intuitively, the equation calculates how many inferred top-N interests are similar to the top-N ground truth. In our study, we set $N = 10$ and use the interests extracted from a user's contributed content as his ground-truth implicit interests.

Then, we randomly partition the ground-truth data of the 25315 employees into ten parts and perform 10-fold cross validation to compute the inference quality. In each round, we leave out the data of one-tenth of the users (testing set), and infer their interests using only the extracted interests for the other nine-tenths of the users (training set). To investigate the effectiveness of combining different types of social content, we conduct the experiment in four conditions: (1) using social bookmarking data only, (2) using file sharing data only, (3) using electronic communication data only, and (4) using all three types of data. Table 1 shows the inference quality in the four conditions.

Condition	Max	Mean	Min	St. Deviation
1	59.4%	19.2%	5.1%	10.7%
2	44.9%	12.7%	3.0%	7.2%
3	62.1%	29.6%	3.8%	14.1%
4	100%	45.1%	4.2%	21.7%

Table 1: The quality of user interest inference.

The comparison results demonstrate the significant advantage of combining multiple sources of social information. The mean quality of combining sources outperform the best mean quality of using one source (electronic communication) by 52%. We attribute the significant improvement to the much wider coverage of the combined social content as illustrated by Figure 1. Specifically, each social media is used for different purpose. For example, social bookmarking is mostly for organizing and sharing web content while electronic communication is used more often for an information worker's daily business. Therefore, the content in the different sources may relate to different aspects of a user's interest. Among the three social media, electronic communication data gives the best performance because of its ubiquity and quantity. Social bookmarking outperforms file sharing, since social bookmarks are more direct and concise indication of user interests [23].

However, Table 1 also shows that the variance of inference quality is huge. We hypothesize that the variance is due to the large variation in people's contributed content and their positions in the information diffusion of the social networks. In Section 6, we test this hypothesis and present a method to predict inference quality. By using the predictor to decide when to infer user interests, applications can improve the mean quality of the results.

5.4 Quality of Inferring Explicit Interests

We also evaluate the effectiveness of inferring explicit user interests from friends. The quality of a user's inferred interests are measured by how much the inferred interests overlap with ground truth. Formally, we measure the quality by precision and recall defined as:

$$\mathtt{Q}_p = \frac{|INF \bigcap GND|}{|INF|}, \quad \mathtt{Q}_r = \frac{|INF \bigcap GND|}{|GND|} \tag{6}$$

where INF is the set of terms with positive weights in the user's inferred explicit interests, GND is the ground truth of the user's explicit interests as specified in his profile.

We then perform 10-fold cross validation to compute the inference precision and recall, using the ground-truth data of the 8005 employees that have both contributed social content and specified explicit interests. We focus on this set of users so that we can study how factors concerning user contributed social content may impact the inference in Section 6. The average inference precision $\mathtt{Q}_p$ is 30.1% with a standard deviation of 26.9%. In contrast, the mean inference recall $\mathtt{Q}_r$ is 61.5% with a standard deviation of 27.6%. For the significantly higher recall, one possible explanation is that a user's interests may often be a small subset of the combined interests of his friends. Therefore, interests inferred from friends can have higher recall. Again, we will show in Section 6 that the inference quality predictor can help applications to improve explicit interest inference.

6. PREDICTING INFERENCE QUALITY

Because there is large variance in the quality of inferring user interests from social neighbors, it is difficult for practical applications to decide whether to utilize the inferred interests. Therefore, it is highly desirable that inference quality can be predicted based on social network features. In this section, we examine a set of relevant network factors and present a regression model to predict the inference quality.

6.1 Factors

We hypothesize that a user's interests are impacted by the type and amount of content contributed in his ego network as well as the structural characteristics of the ego network. Specifically, we examine five factors including *user activeness* measured by the amount of contributed content, *network in-degree*, *network out-degree*, *between centrality*, and *user management role* in the organization. We focus on studying the factors in the user's three-degree ego network, where the most social influence comes from. For each factor, we extract corresponding feature for three sub ego networks including one-degree neighbors, two-degree neighbors, and three-degree neighbors respectively. This allows us to assess the different influence from neighbors of different degrees of separation. In addition, the role of the user himself in the social networks may influence interest inference. For example, a user that plays an important role in the information flow within the network may be more likely to share interests with his neighbors. Therefore, we also extract ego network feature of the user himself including in-degree, out-degree, betweenness and user management role.

6.1.1 User activeness

We measure the activeness of user i in the network using the amount of content he contributed. Let $a(i)$ be a (3×1) vector $\langle a_0(i), a_1(i), a_2(i)\rangle$, where $a_0(i)$, $a_1(i)$, $a_2(i)$ are the amount of the user's contribution in social bookmarking, file sharing, and electronic communication, respectively. Their values are normalized to $[0, 1]$ by dividing by the maximum contribution in each type of media.

The social influence user i receives on a subject could be mostly from a highly active friend in the user's ego network, or the collective influence from all neighbors is more important. To distinguish the two situations and compare their relative importance, we calculate both the sum and the maximum value of the activeness in the ego network of user i. Moreover, each neighbor's activeness should be weighted by his social influence on user i:

$$A_j(i) = \sum_{v \in V_k} [w_{vi} \cdot a_j(v)], \quad A'_j(i) = \max_{v \in V_k} [w_{vi} \cdot a_j(v)] \quad (7)$$

where w_{vi} is defined in Equation 3 to indicate the social influence user v has on user i, $j = 0, 1, 2$ denotes the type of the social media, and V_k is the node set of the sub ego network containing neighbors of degree k.

6.1.2 Network in-degree and out-degree

Network in-degree and out-degree measure the number of electronic communication (e.g., email) between a user and other users in a network. For user i, in-degree is measured as the number of email sent to i, while out-degree is measured as the number of emails sent out from i. For a sub ego network V_k of user i, we also calculate both the sum and the maximum value of the in-degree as:

$$I(i) = \sum_{v \in V_k} [w_{vi} \cdot inDeg(v)], \; I'(i) = \max_{v \in V_k} [w_{vi} \cdot inDeg(v)]$$

where V_k is the sub ego network containing neighbors of degree k, $inDeg(v)$ is the number of emails coming into user v. Similarly, we calculate the sum and the maximum value of the out-degree as:

$$O(i) = \sum_{v \in V_k} [w_{vi} \cdot outDeg(v)], \; O'(i) = \max_{v \in V_k} [w_{vi} \cdot outDeg(v)]$$

where $outDeg(v)$ is the number of emails coming out from user v. In addition, we include the ego attributes $inDeg(i)$ and $outDeg(i)$ as features for user i.

6.1.3 Betweenness centrality

For an individual node i in social networks, the betweenness centrality $b(i)$ measures the relative importance of the node in the information flow within the network [9]. Hence, a user with high betweenness value might have high influence on his neighbors. Specifically, $b(i)$ is defined as the probability that node i will fall on the shortest path between any two other individuals in a network:

$$b(i) = \sum_{l<m} [\eth_{lm}(i)/\eth_{lm}] \quad (8)$$

where $\eth_{lm}(i)$ is the number of shortest geodesic paths from l to m that pass through a node i, and $\eth_{lm}$ is the total number of shortest geodesic paths from l to m. For the sub ego network V_k of user i, we again compute the sum and the maximum value of betweenness centrality as:

$$B(i) = \sum_{v \in V_k} [w_{vi} \cdot b(v)], \qquad B'(i) = \max_{v \in V_k} [w_{vi} \cdot b(v)] \quad (9)$$

where V_k is user i's sub ego network containing neighbors of degree k. Finally, we include the ego betweenness value $b(i)$ as a feature for user i.

6.1.4 User management role

A user's formal role in the organization (e.g., manager) may also impact his influence to his friends. Therefore, we include a factor on user formal role. For user i, we define $m(i)$ to describe his formal role with 3 possible values: 0 (non-managers), 1 (line managers), and 2 (executives). In our data set, there are 81.3% non-managers, 15.0% managers and 3.7% executives. In addition, for the sub ego network V_k of user i, we compute the sum ($M(i)$) and the maximum value ($M'(i)$) of user management role as:

$$M(i) = \sum_{v \in V_k} [w_{vi} \cdot m(v)], \quad M'(i) = \max_{v \in V_k} [w_{vi} \cdot m(v)] \quad (10)$$

6.2 Predicting By Regression

Regression is a classic statistical problem which tries to determine the relationship between two random variables $x = (x_1, x_2, ..., x_F)$ and y. In our scenario, dependent variable y can be the inference quality C, and independent variable x can be a feature vector based on the factors described in Section 6.1. Specifically, for user i,

$$x = \langle x'_k, inDeg(i), outDeg(i), b(i), m(i)\rangle, (k = 1, 2, 3) \quad (11)$$

where $inDeg(i)$, $outDeg(i)$, $b(i)$ and $m(i)$ are ego features for user i, and x'_k is a (14×1) feature vector for the sub ego network V_k with neighbors of degree k:

$$x'_k = \langle A_0, A_1, A_2, A'_0, A'_1, A'_2, I, I', O, O', B, B', M, M'\rangle \quad (12)$$

Here x'_k includes the sum and maximum values of user activeness in different types of social media, network in-degree and out-degree, betweenness, and user management role. All feature values are normalized to $[0, 1]$.

Given the features, we use support vector regression (SVR) predict inference quality. In SVR, the input x is first mapped onto a high dimensional feature space using a nonlinear mapping, and then a linear model is constructed in this feature space. SVR uses a so called ε-insensitive loss function:

$$L_\varepsilon = \begin{cases} 0 & \texttt{if } |y - f_\omega(x)| < \varepsilon \\ |y - f_\omega(x)| & \texttt{otherwise} \end{cases}$$

where ε is a predefined deviation threshold, and $f_\omega(x)$ is the regression function to predict y which has a parameter ω. Then the regression is formalized as the following problem:

$$\texttt{minimize} \quad \tfrac{1}{2}\|\omega\|^2 + \mathbb{H}\textstyle\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$
$$\texttt{subject to} \begin{cases} y_i - f_\omega(x) \le \varepsilon + \xi_i \\ f_\omega(x) - y_i \le \varepsilon + \xi_i^* \\ \xi_i \,,\, \xi_i^* \ge 0 \end{cases}$$

where $\mathbb{H}$ is a constant, and ξ_i, ξ_i^* $(i = 1, ..., l)$ are slack variables introduced for the optimization to measure the deviation of training samples outside ε insensitive zone.

In our study, we collect ground-truth interest data of the 25315 contributors and their ego network features to build

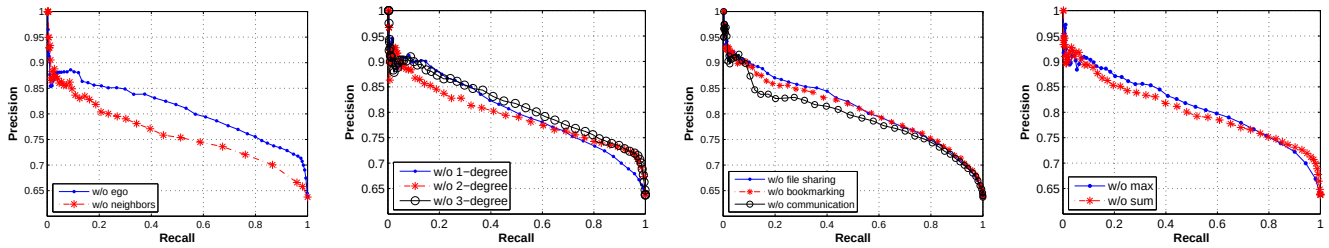

Figure 3: Compare the impact of different features on prediction: (a) ego *vs.* neighbors, (b) degrees of separation, (c) types of social content, and (d) maximum and sum values of the features.

SVR model. We use the support vector regression implementation in SVM-Light [12]. In our support vector regression experiments, we use sigmoid kernel function $\tanh(\mathfrak{s} \cdot x_i^T \cdot x_j + \mathfrak{c})$ with parameter $\mathfrak{s} = 0.2$. Other parameters such as ε are set to default. Next, we evaluate the effectiveness of the regression in Section 6.3 and 6.4.

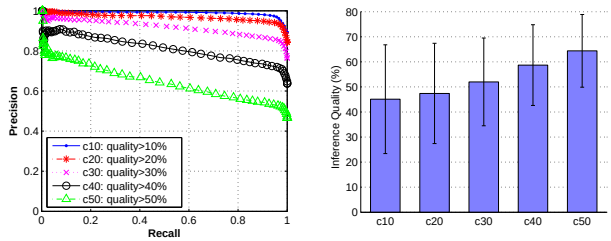

Figure 2: (a) The precision-recall curves for implicit interest inference quality classification. (b) The quality improvements after using inference quality classification.

6.3 Results on Implicit Interest Inference

To evaluate the quality of the inferred top-N implicit interests, we use the prediction to classify "high quality" inferred interests (e.g., quality $C > 50\%$). In practice, the criteria for high quality can vary according to application scenarios. To classify "accurate" interest inference, we test whether the quality prediction is larger than a threshold TH. The precision and recall for a particular TH can be defined as:

$$\texttt{precision}_{TH} = \frac{|E_H \bigcap E_{TH}|}{|E_{TH}|}, \quad \texttt{recall}_{TH} = \frac{|E_H \bigcap E_{TH}|}{|E_H|}$$

where E_H is the ground truth set of "accurate" inferences that satisfy our preselected quality criteria, and E_{TH} is the set of predictions larger than TH. A precision-recall curve can be derived by varying the threshold TH.

We randomly partition the ground-truth user interest data into ten parts and use 10-fold cross validation to evaluate the classification performance based on support vector regressions. Five criteria are used to classify the inference quality. The resulting precision-recall curves are shown in Figure 2(a). We observe that satisfactory performance can be achieved by our quality prediction method. For example, to classify inferred interests that have quality larger than 50%, our method gives a precision of 62%, at a recall of 50%. Therefore, we can use this model to inform search and recommender applications when to leverage the inferred interests. In that situation, we only infer a user's interests from his friends when the classifier predicts the result will be better than the predefined criteria. Figure 2(b) shows the mean and standard deviation of the inference quality when using the classifier. Compared to the results in Table 1, we can observe that the mean quality is significantly improved, while the standard deviation is reduced. For example, using the c50 classifier, the mean quality is improved from the 45.1% in Table 1 to 64.4%, an improvement of 42.8%.

6.3.1 *Comparing Different Features*

To understand the relative impact of various features used in predicting interest inference quality, we perform "leave-one-feature-out" comparisons. That is, to compare the impact of two features f_1 and f_2, we perform two 10-fold cross validation as described earlier. For the first time, we use all features but f_1. In the second experiment, we leave out f_2. After that, we compare the precision-recall curves from the two experiments to understand which feature has bigger impact on prediction. In our study, we may also leave a set of features out to examine their impact on prediction. In all comparisons, we use the c40 criteria ($quality > 40\%$).

First, we compare the impact of the features describing ego ($inDeg(i)$, $outDeg(i)$, $b(i)$, and $m(i)$ in Equation 11) and the features describing neighbors (x'_k in Equation 11). Figure 3(a) shows the corresponding two precision-recall curves. We observe that it is possible to predict inference quality only using features about ego role in the information flow within the network (the red dashdot curve with star marker). However, the performance of ignoring ego features is mostly better than leaving out features about neighbors. That is, the features on neighbors are more important.

Second, we compare the features on neighbors of different degrees of separation. In Figure 3(b), the precision-recall curve obtained by ignoring three-degree neighbors has the best overall performance. This agrees with previous studies that social influence decreases quickly as the degree of separation increases. However, to our surprise, leaving out one-degree neighbors has smaller impact than two-degree neighbors for recall between $[0.1, 0.6]$. Further investigation into our data set gives a possible explanation. There are on average 21 times more two-degree neighbor than one-degree neighbors. Thus the two-degree neighbors can provide much more information than one-degree neighbors. Therefore, the impact of neighbors of different degrees of separation is the

result of balancing the two factors: the amount of information provided and the degree of separation.

Next, the impact of the three types of social information sources is compared. In our study, we have focused on three types of media: social bookmarking, file sharing and electronic communication (e.g., email and instant messaging). Figure 3(c) illustrates the three precision-recall curves. The results show that ignoring electronic communication has the largest impact on prediction. This is understandable since the amount of electronic communication data is much larger: 20 million compared to less than 1 million in other two types of media. Besides electronic communication, social bookmarking has a bigger impact than file sharing. This can be explained by the fact that social bookmarks are usually meaningful tags used to improve information search and organization [18, 11, 3]. Therefore, it is easier to extract user interests from social bookmarks than shared files.

Moreover, we compare the impact of the maximum and sum values of the factors, which are both used as the features to predict quality (see Equation 12). For a user, the maximum values of the factors correspond to the most prominent neighbor (e.g., the most active neighbor). In contrast, the sum values indicate the collective influence of all the neighbors. Figure 3(d) shows the precision-recall curves. We observe that leaving out the sum values has a bigger overall influence on the classification performance. Therefore, the collective influence from all neighbors is still noticeably more important than the influence from a single most prominent friend. On the other hand, the features on the most prominent friend can already achieve a reasonable performance (e.g., $< 5\%$ difference in performance), especially in high-recall, low-precision situations.

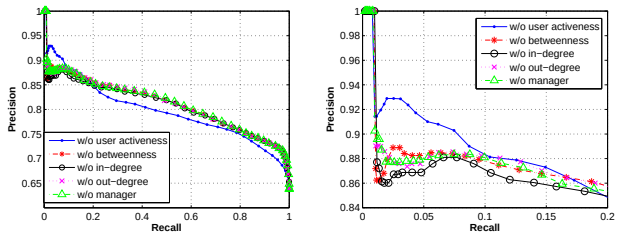

Figure 4: Compare the impact of five factors including user activeness, network in-degree, out-degree, betweenness and user management role. (a) An overall view, (b) a detailed view with $0 \leq recall \leq 0.2$.

In addition, we assess the relative impact of the five factors: *user activeness, network in-degree, network out-degree*, *betweenness centrality*, and *user management role*. The five corresponding precision-recall curves are illustrated in Figure 4. In Figure 4(a), we observe that overall *user activeness* has more significant impact than the other four factors. In contrast, a more detailed comparison in Figure 4(b) shows that *network in-degree, out-degree, betweenness* and *user management role* have bigger impact in the scenario with high precision (precision $> 85\%$) and low recall (recall $< 20\%$). Among the social network features, network in-degree has the biggest impact. This indicates that it may be accurate to infer an inactive user's interests if he receives information from many active friends. The implication is that applications that demands high precision of prediction should rely more on social network structure properties like network in-degree.

Finally, we examine the impact of different user management role (e.g., non-managers, line managers and executives). For each role, we randomly leave out 253 users (i.e., 1% of the user population in our study) with that role, when inferring user implicit interests using Equation 2. We repeat the experiment 10 times for each management role. The mean inference accuracies for leaving out each management role are: 42.7% (non-managers), 42.1% (line managers) and 40.2% (executives). Although the differences are not statistically significant, the trend agrees with the intuition that higher level managers have more social influence.

6.4 Results on Explicit Interest Inference

To verify our findings against explicit user interest inference, we apply the regression model trained by the implicit user interests data in Section 6.3 to predict the inference quality. We first choose the recall of the inferred results $\mathtt{Q}_r$ (Equation 6) as the inference quality measure, since it is shown to have higher performance (Section 5.4). Figure 5 illustrates the precision-recall curves of the explicit interest inference quality classification using three different quality criteria. For example, "Criterion: 60%" means that $\mathtt{Q}_r > 60\%$. Again, we can use the classification to decide when to perform interest inference. Using the "Criterion: 70%" classifier, the mean $\mathtt{Q}_r$ is improved from 61.5% to 85.7% (a 39.3% improvement), the standard deviation is reduced from 27.6% to 17.4%. Similarly, we also obtain improvements on the precision of the inferred results $\mathtt{Q}_p$. The mean $\mathtt{Q}_p$ is improved from 30.1% to 60.5% (a 101% improvement), and the standard deviation is reduced from 26.9% to 21.3%. In addition, the comparison of different factors' impact gives similar trends observed in Figure 3 and 4.

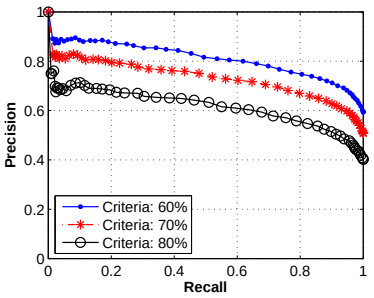

Figure 5: The precision-recall curves for explicit interest inference quality classification.

7. DISCUSSION

We studied the quality of inferring user interests from social neighborhood. Our findings demonstrate that the performance can be promising but the variance is large. This has implications for the search and recommendation applications that rely on user interest models inferred from social neighbors to provide personalized results. For example, a social blogging site may try to provide personalized blog recommendation to inactive users, which can encourage them to participate more actively. In that situation, the application needs to infer users' interests from their friends. It can

use the predictor to decide when to infer user interests from friends and achieve reasonable quality (e.g., mean accuracy 64.4%). In addition, the applications need combine all available social content sources to increase the inference quality. For example, sources that provide larger amount of information (e.g., email) or are easier to extract interests (e.g., social bookmarking) can be given more weights. Furthermore, neighbors with one or two degrees of separation have the most impact on inferring one's interests. Two-degree neighbors may even have a bigger impact since there are much more of them and yet not too far away. Finally, users' roles in formal networks also impact the inference. For instance, higher level managers can be more influential.

Our findings are based on social networks inferred from people's electronic communication including email and instant messaging. In practice, the findings may also be applied to social networks constructed from other types of interaction among people, depending on their availability. For example, the strength of the tie between two people can be inferred from their activities on online social sites (e.g., blogs, forum), co-authorship of papers [24], or face-to-face interaction captured by physical sensors [28].

8. CONCLUSION

In this paper, we present a study on the quality of inferring user interests from friends in one of the largest global organizations. We demonstrate that there exist large variance of the inference quality when user contributed content considerably vary and the content types are diverse. To allow search and recommendation applications make informed decisions on when to utilize inferred user interests, we further investigated relevant factors and present a method to predict inference quality based on network features including user activeness, network in-degree, out-degree, betweenness centrality and user management role. Our experiments validate the effectiveness of the prediction method and compare the relevant importance of the factors. Our findings can be useful for social applications with widely varied participation rate so that the interests of many people can only be inferred from their friends. In particular, our results can be leveraged to provide new users with personalized recommendations, especially in our system where a user's social networks can be automatically constructed based on archived years of communications. This can motivate the new users to actively participate in social applications [6].

We are planning to incorporate the inferred user interests to provide better personalized results in enterprise expert finding and content recommendation systems [16]. In the future, we plan to examine additional network properties that may affect social influence, such as the *network backbone* structure recently observed in email communication networks [13]. Another future direction is to study how well the dynamic evolution of users' interests can be inferred from friends. Such study can increase our understanding on the temporal aspect of user interest models in order to provide up-to-date personalized search and recommendations.

9. ACKNOWLEDGEMENT

This research is continuing through participation in the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under Agreement Number W911NF-09-2-0053.

10. REFERENCES

[1] E. Agichtein, E. Brill, and S. T. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR*, pages 19–26, 2006.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD*, pages 7–15, 2008.

[3] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *CIKM*, pages 193–202, 2008.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] M. Brzozowski, T. Sandholm, and T. Hogg. Effects of feedback and peer pressure on contributions to enterprise social media. In *CSCW*, pages 61–709, 2009.

[6] M. Burke, C. Marlow, and T. M. Lento. Feed me: motivating newcomer contribution in social network sites. In *CHI*, pages 945–954, 2009.

[7] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007.

[8] D. J. Crandall, D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD*, pages 160–168, 2008.

[9] L. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[10] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

[11] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *WSDM*, pages 195–206, 2008.

[12] T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. B. Schölkopf and C. Burges and A. Smola. (ed.). The MIT Press, 1999.

[13] G. Kossinets, J. M. Kleinberg, and D. J. Watts. The structure of information pathways in a social communication network. In *KDD*, pages 435–443, 2008.

[14] R. T. Leenders. Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24:21–47, 2002.

[15] D. Liben-Nowell and J. M. Kleinberg. The link prediction problem for social networks. In *CIKM*, pages 556–559, 2003.

[16] C. Lin, K. Ehrlich, V. Griffiths-Fisher, and C. Desforges. Smallblue: People mining for expertise search,. *IEEE Multimedia Magazine*, 15(1):78–84, 2008.

[17] H. Ma, H. Yang, M. R. Lyu, and I. King. SoRec: social recommendation using probabilistic matrix factorization. In *CIKM*, pages 931–940, 2008.

[18] D. R. Millen, J. Feinberg, and B. Kerr. Dogear: Social bookmarking in the enterprise. In *CHI*, pages 111–120, 2006.

[19] B. Piwowarski and H. Zaragoza. Predictive user click models based on click-through history. In *CIKM*, pages 175–182, 2007.

[20] A. Shepitsen, J. Gemmell, B. Mobasher, and R. D. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys*, pages 259–266, 2008.

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

[22] X. Song, B. Tseng, C. Lin, and M.-T. Sun. Expertisenet: Relational and evolutionary expert modeling. In *Proc. of Intl. Conf. on User Modeling*, pages 99–108, 2005.

[23] J. Stoyanovich, S. A. Yahia, C. Marlow, and C. Yu. Leveraging tagging to model user interests in del.icio.us. In *AAAI Spring Symposium on Social Information Processing*, 2008.

[24] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD*, pages 807–816, 2009.

[25] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR*, pages 449–456, 2005.

[26] R. W. White, P. Bailey, and L. Chen. Predicting user interests from contextual information. In *SIGIR*, pages 363–370, 2009.

[27] L. Wu, C. Lin, S. Aral, and E. Brynjolfsson. Value of social network – a large-scale analysis on network structure impact to financial revenue of information technology consultants. In *The Winter Conference on Business Intelligence*, 2009.

[28] L. Wu, B. Waber, S. Aral, E. Brynjolfsson, and A. Pentland. Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task. In *International Conference on Information Systems*, 2008.