

# When *Diversity* Meets *Speciality*: Friend Recommendation in Online Social Networks

Hao Wu<sup>†</sup>, Vikram Sorathia, Viktor K. Prasanna

<sup>†</sup>Department of Computer Science

Ming Hsieh Dept. of Electrical Engineering

University of Southern California

Email: {<sup>†</sup>hwu732, vsorathi, prasanna}@usc.edu

## ABSTRACT

Online social networks improve social experience by connecting users with common interests. Similar to real life, seeking good friends is much easier with recommendations in online social networks. In this paper, we investigate a series of problems related to friendship formation in the hope of improving friend recommendation in social networks. Specially, we seek to understand whether users who contribute more are more popular among other users, whether users like to make friends with popular users and the role difference of users with different diversity of individual interests in friendship formation. We propose a novel approach based on topic modeling to characterize the interest diversity degree of each user. The interest diversity features are used to help predict friend relationships between users. The experimental results on three large-scale datasets demonstrate the effectiveness of our method.

## I INTRODUCTION

Online Social network services have thrived in great popularity recent years. Famous websites such as *Facebook*<sup>1</sup>, *Twitter*<sup>2</sup>, *Flickr*<sup>3</sup>, *Last.fm*<sup>4</sup>, *De.licio.us*<sup>5</sup> have attracted tremendous numbers of users and play an increasing important role in online interaction. By connecting users with similar professional background or common interests, online social networks open up a new channel for information sharing and social networking. The open-ended nature of their applications motivates rich user-generated content, including tags, text document, multimedia, and so on [14].

One fundamental phenomenon in social network services is friendship formation. Members (users) make friends with each other through social interactions and information exchange. Analysis of friendship formation in online social networks help understand many sociological and psychological problems such as community gener-

ation [1, 15], interest identification and opinion forming [5]. It can also greatly facilitate direct member matching or friend recommendation. The significance of member matching is two-sided [4, 18]. In the macro scope, matching members is critical for the initial growth and further development of the online social network. In the micro level, a member in online social network may be frustrated to find good friends from a tremendous number of irrelevant users. Suggesting relevant users with common interests to each individual can help improve user experience. Most social network websites match members based on the number of mutual friends. This method suffers the drawback of interest mismatch and it is useless to expand the circle of the members, because someone who has many common friends with you probably already known to you.

The social network built on member friendships can be naturally modeled as a graph, where each node represents a member, and edges model the friend links. Analyzing the proximity of two members for member matching is fundamentally related to link prediction problem [10, 12, 17], which predict the edges that will be added to the network in the near future given a present snapshot of a social network. But this line of work mainly focus on the network structure and its evolution, while the intrinsic properties of the nodes (users) in the network are ignored. A sophisticated friend recommendation systems should consider the social interactions between users which help build the friendship, and the individual interests of users as well.

The rich user-generated content in online social networks impose challenges to mine user interests and regular behaviors for recommendation [11, 13]. The information of user behavior is often scattered in both social links and content reflecting user interests such as self-generated profile, semantic tagging, browsing action, interaction with other members and so on. Recent work [14] focused on the usage of shared tags with the existing so-

<sup>1</sup><http://www.facebook.com/>

<sup>2</sup><http://twitter.com/>

<sup>3</sup><http://www.flickr.com/>

<sup>4</sup><http://www.last.fm>

<sup>5</sup><http://www.delicious.com/>

Table 1: Statistics of the Datasets

dataset	genre	users	items	tags	friend edges
<b>Last.fm</b>	music	99,405	1,393,559	281,818	3,151,283
<b>Flickr</b>	photo	319,686	28,153,045	1,607,879	NA
<b>Del.ious.us</b>	web	532,924	17,262,480	2,481,698	NA

cial network for link prediction, but the interplay of the user social interactions and friendships is never captured in the rich context of online social networks. For example, no previous work analyzes the diverseness of individual tastes reflected by the social interactions, which is one of the most influential factor in recommendations [6].

## 1 CONTRIBUTIONS

In this paper, we take advantage of the social interactions that reflect the user interests, in order to predict friendship links for recommendation. Our work brings sociological and psychological insights into friendship formation problem in online social networks. The novelty and main contributions can be summarized as:

1. We investigate a series of questions associated with friendship formation. Specially, whether those users contribute more annotations are more popular, whether users like to make friends with popular users, and most importantly how does users with different diverseness of individual tastes account for friendship formation?
2. We propose a novel diverseness measurement of individual interests based on topic modeling. The diverseness degree of user tastes are represented as the semantic diverseness of the items annotated by each user. We use the collective knowledge of crowd tagging as the resource to identify such diverseness.
3. We adopt the diversity features in learning algorithm to predict the friendship. The results show the robustness of the diversity features. When addressing the friendship prediction of users with different diverseness of individual interests, we have interesting observations that users with high-degree diversity of interests are more likely to form friendship with each other, and the friendships are more predictable.

<sup>6</sup><http://en.wikipedia.org/wiki/Last.fm>

## II METHODS AND EXPERIMENTS

### 1 DATASETS

We investigate three popular social networks across music listening (*Last.fm*), photo sharing (*Flickr*), and web bookmarks (*Del.icio.us*). The statistics of the three datasets are listed in Table 1.

We mainly focus on a large-scale dataset of *Last.fm*, where the friendships are available, and the statistics of the dataset is illustrated in Table 2. We both conduct the diverseness analysis of individual tastes as well as the friendship prediction task in *Last.fm*. For *Flickr* and *Del.icio.us* datasets where the user friendships are not completely visible, we mainly focus on the analysis of the diverseness of individual interests. The main user activity in these online social networks is annotation which can be represented as tuples (*user, item, tag*). The annotations are mainly user-generated in online social networks.

There are various types of social interactions going on in online social media. We illustrate different user activities and interactions in online social networks in Fig. 1. Typically, we take *Last.fm* for example of introduction. *Last.fm* is featured by a music system called “Audioscrobbler” that can suggest new music to each user tailored to the user’s own preferences. Users can listen to their personal music collection, listen to internet radio services, label the music tracks, artists or albums with tags, etc<sup>6</sup>. User profile is thus built with user name, avatar, date of registration and total number of tracks played, as well as the following aspects of user interactions or activities.

- **Friends** can be added by registered members if they have similar tastes of music or shared groups. Similar to any other online social networks. The friendship represents a strong relationship of common interest. The friendship in *Last.fm* is mutual, i.e., if A is a friend of B.
- **Items** are related to services such as music tracks, artists, albums and radios, etc. Users can tag them with words.

Table 2: User properties of the *Last.fm* dataset

property	min	median	mean	std	max
# of friends	1	2	6.5	42.0	19,412
# of groups	1	6	13.6	40.9	3,737
# of items	1	8	84.9	506.6	48,830
# of distinct tags	1	7	28.2	97.8	4,666
# of tags	1	15	208.5	1,545.8	172,448

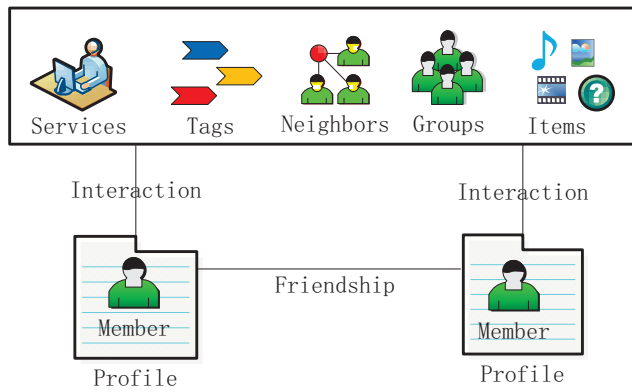


Figure 1: Data graph in online social networks

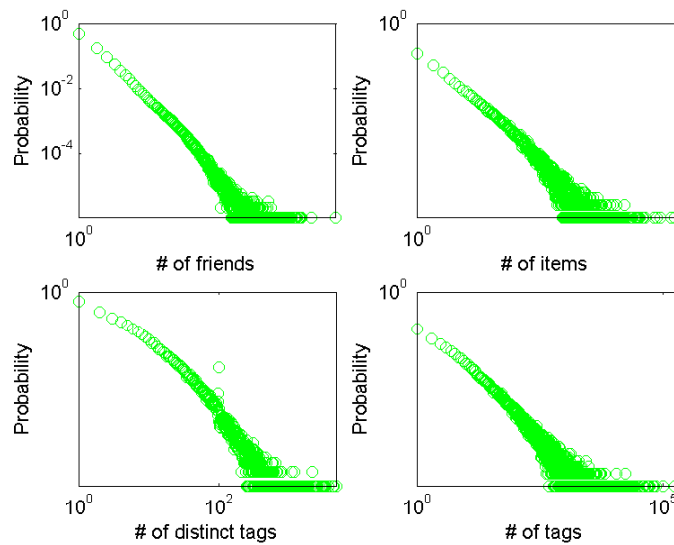


Figure 2: Distributions of the number of friends, items, distinct tags and tag assignments per user in *Last.fm*.

- **Tags** are supported by online social networks such as *Last.fm* for user-end labeling of tracks, artists, albums, and radios to create a site-wide folksonomy of music. This tagging can be by genre, mood, artist characteristic, or any other form of user-defined classification, which can facilitate user to browse.
- **Groups** are formed by users with common interest. A typical group may be created by fans of an artist or a genre of music. Most groups are open to all users to join, but may need approval of membership.

The distributions of the number of friends, items, distinct tags and tag assignments for each user in *Last.fm* dataset are illustrated in Fig. 2. Similar to some other online social networks [9], we observe long tail distributions of the different user interactions (with other users, items, tags) from the curves. We have similar observations in *Flickr* and *Del.ious.us*.

## 2 CONTRIBUTION VS. POPULARITY

We then investigate how the *contribution* is related to the *popularity* of a user. For *Last.fm* dataset, We plot the correlations between number of items annotated and number of friends each user have. As depicted in Fig. 3(a), we observe there is no explicit pattern reflecting the correlations. We then calculate the mean number of items annotated by the users with specified number of friends, as shown in Fig. 3(b). We can see the individual variance of popularity masks a pronounced effect of contribution on mean. Users with larger number of friends on average annotate more items. which is illustrated by the left part of the figure. However, the scatter pattern of the right part of Fig. 3(b) suggests it is not absolutely true as the number of a user's friends has exceeded some certain value. The popularity of these users with very large number of friends is not directly related to their contributions.

## 3 ECCENTRICITY OF MAKING FRIENDS

Next, we try to understand in general whether users like to make friends with popular users or unpopular users. To characterize the variance in individual preferences of making popular friends, we first rank all users by the number of friends each individual has, which is the *popularity* rank of each user. We then define a measure called *eccentricity* for each user, which is the median *popularity* rank of all the friends the user has. In particular, higher eccentricity corresponds to on average make less popular friends. Fig. 4 depicts the distribution of user eccentricity in *Last.fm*. We can observe there is significant variation

between individuals, demonstrated by relatively wide interquartile ranges of the eccentricity. The distribution of user eccentricity exhibits a heavy tail pattern. In general, a large proportion of users make friend who are popular, while other users exhibit relative eccentricity of making unpopular friends.

## 4 INTEREST DIVERSITY

It is shown that users with *diverse* tastes or *special* tastes have different reactions to recommendations [6]. It is interesting to investigate the diverseness of individual interests in online social networks, which can help build friendship recommendation and item recommendation systems. Recommendation strategy can be various regarding to users with different levels of interest diversity. Intuitively, a user with diverse interests will easily adopt recommendations, especially the popular ones (users or items). For example, suppose there is a user that has broad interests in comedy, cartoon, scientific, classic movies, etc. He or she probably like to adopt any movie with high-quality and like to make friends also with broad interests. So an straightforward strategy could be recommending this user with recent popular movies or users who are popular. However, other users may have propensity toward some special genres. Those users account for the long tail of product consumption [6]. For examples, for someone who loves watching old times classic movies or whose tastes only lie on cartoon movies, matching their special interests will be important in recommendations. The first step is to measure the diversity degree of each user. We propose a novel approach based on topic modeling. It is based on the intuition that the diversity of semantic annotation of items that a user has interacted with reflects the interest diversity of the user. To represent the semantic diversity of each item, we leverage the collective wisdom of crowd tagging. In particular, each item is regarded as a *document* with *word* representation of its tags annotated by the crowd of users. We use topic modeling method, Latent Dirichlet Allocation (LDA [3]) to model the co-occurrence of *items* and *tags*. In our method, LDA gives the following generative process for tags  $\mathbf{x}$  that are used by the crowd of users to annotate an item  $i$ .

1. Choose topic distribution  $\Theta_i \sim \text{Dirichlet}(\alpha)$
2. For each of the  $N$  tags,  $x_n \in \mathbf{x}$ 
  - (a) Choose a latent topic  $z_t \sim \text{Multinomial}(\Theta_i)$
  - (b) Choose a tag  $x_n$  from  $p(x_n|z_t, \beta)$ , a multinomial probability conditioned on the topic  $z_t$ .

This topic modeling procedure is illustrated as in Fig. 5, where we use music sharing community (e.g., *Last.fm*) as











