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

Hao Wu†, Vikram Sorathia, Viktor K. Prasanna
†Department of Computer Science
Ming Hsieh Dept. of Electrical Engineering
University of Southern California
Email: {†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*[1], *Twitter*[2], *Flickr*[3], *Last.fm*[4], *De.licio.us*[5] 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 generation [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-

[1] http://www.facebook.com/
[2] http://twitter.com/
[3] http://www.flickr.com/
[4] http://www.last.fm
[5] http://www.delicious.com/

ISBN: 978-1-62561-004-1

Table 1: Statistics of the Datasets

dataset	genre	users	items	tags	friend edges
Last.fm	music	99,405	1,393,559	281,818	3,151,283
Flickr	photo	319,686	28,153,045	1,607,879	NA
Del.ious.us	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.

[6] http://en.wikipedia.org/wiki/Last.fm

II METHODS AND EXPERIMENTS

1 DATESETS

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[6]. 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.

ISBN: 978-1-62561-004-1

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

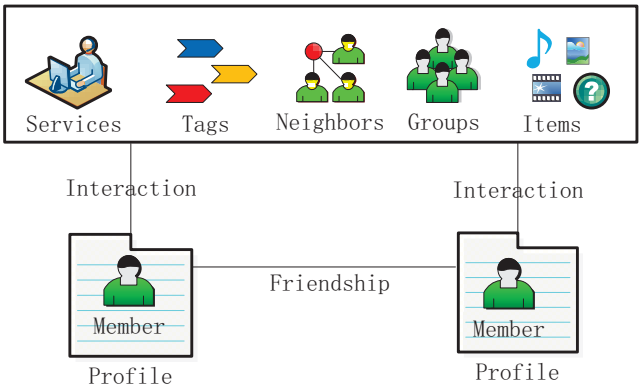


Figure 1: Data graph in online social networks

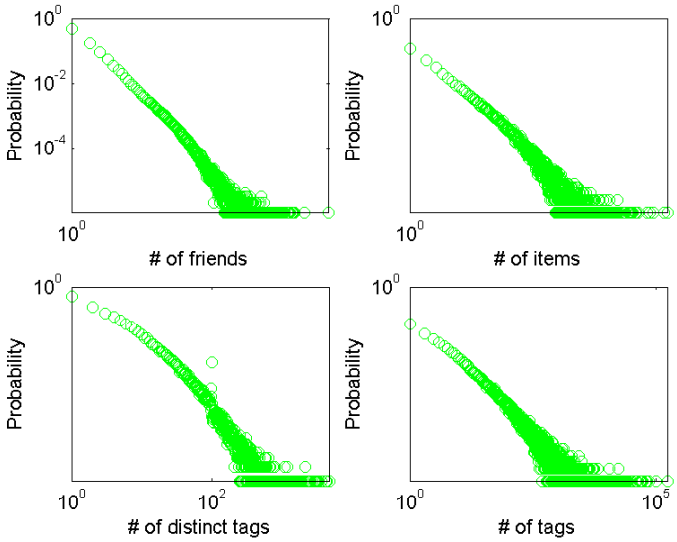


Figure 2: Distributions of the number of friends, items, distinct tags and tag assignments per user in *Last.fm*.

- **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.

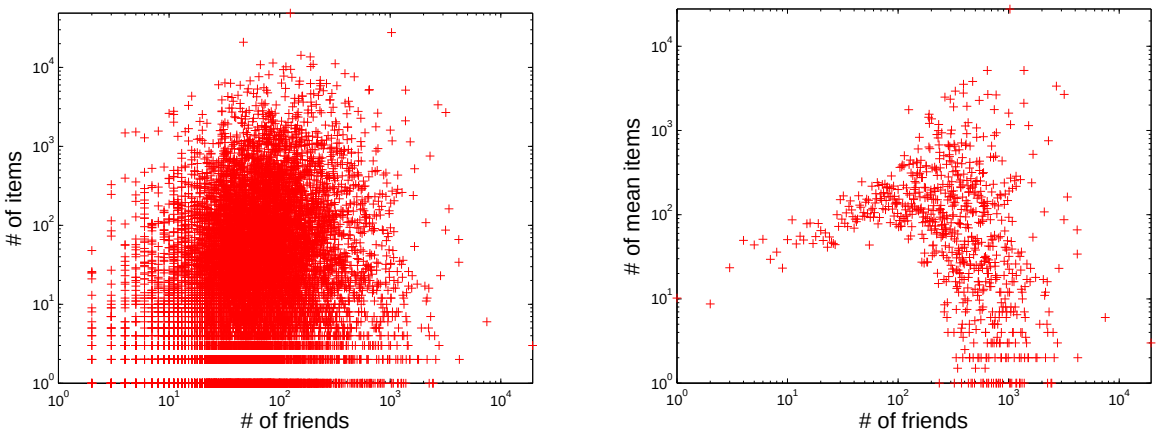
4 INTEREST DIVERSITY

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 $\mathbf{\Theta}_i \sim$ Dirichlet (α)
2. For each of the N tags, $x_n \in \mathbf{x}$
   (a) Choose a latent topic $z_t \sim$ Multinomial ($\mathbf{\Theta}_i$)
   (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

ISBN: 978-1-62561-004-1



(a) contribution (b) mean contribution

Figure 3: Contribution (# of tagged items) VS. Popularity (# of friends) in *Last.fm*

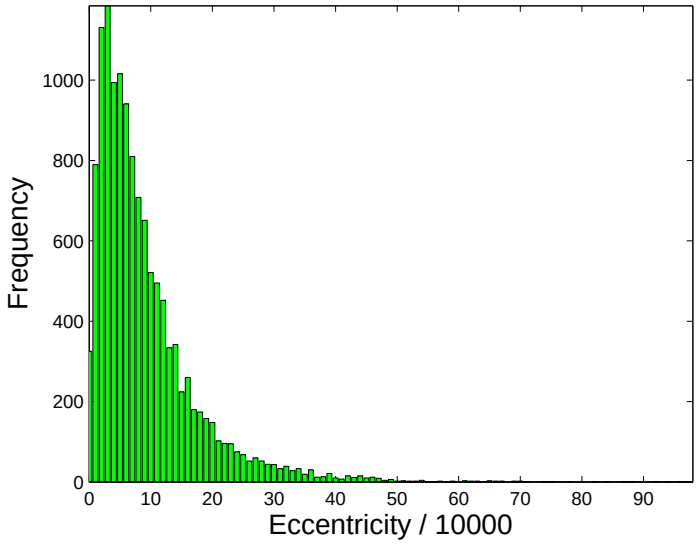


Figure 4: Distribution of Eccentricity in *Last.fm*

Table 3: Topic modeling results for *Del.icio.us* where $T = 100$

TOPIC1	TOPIC2	TOPIC3	TOPIC4	TOPIC5
php 0.27	games 0.33	personal 0.09	news 0.52	business 0.22
comics 0.16	game 0.05	imported 0.04	daily 0.07	marketing 0.08
programming 0.06	gaming 0.04	realestate 0.03	media 0.07	advertising 0.05
webdev 0.06	fun 0.04	test 0.03	magazine 0.06	management 0.03
webcomics 0.02	juegos 0.03	pessoal 0.03	newspaper 0.02	ideas 0.02
mysql 0.02	rpg 0.02	housing 0.02	journalism 0.01	startup 0.02
comic 0.02	retro 0.02	zope 0.02	german 0.01	help 0.02
coding 0.02	videogames 0.01	ingenieria 0.01	newspapers 0.00	ecommerce 0.01
scripts 0.02	emulation 0.01	adsl 0.01	magazines 0.00	communication 0.01
framework 0.01	online 0.01	five-dollarshake 0.01	nachrichten 0.00	entrepreneurship 0.01

ISBN: 978-1-62561-004-1

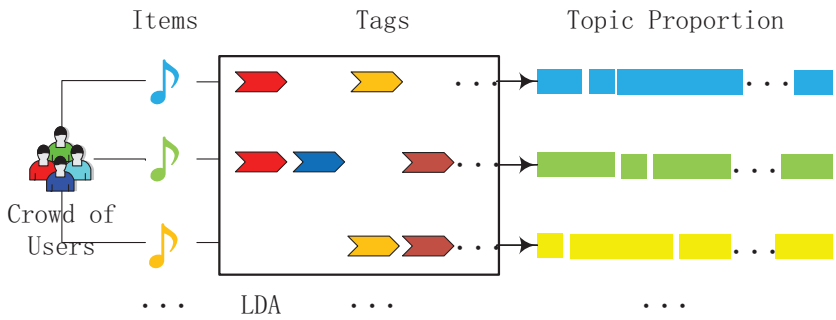


Figure 5: Topic Modeling for Interest Diversity

an example. This method largely reduces the complexity of analysis the content of each item directly [16]. For example, each music has complicate audio features and each photo has high-dimensional pixel features, which causes the complexity to use the content features. The resulting T-dimensional topic proportion of each item i after topic modeling $\mathbf{\Theta}_i = [\theta_i^1, \theta_i^2, ..., \theta_i^T]^\top$ is used to characterize the semantic categories of the item in our method. We then calculate the variance of each topic dimension t of all the I items annotated by a user $\mathbf{\Theta}^t = [\theta_1^t, \theta_2^t, ..., \theta_I^t]^\top$. The mean value of the variances of all topic dimension is used to measure the diverseness of each corresponding user. Eq. 1 illustrates the calculation,

$$DiversityScore = \sum\nolimits_t var(\mathbf{\Theta}^t)/T \qquad (1)$$

We present some illustrative examples of the topic modeling results using Gibbs sampling [7]. Table 3 show the results of topic modeling using LDA with 100 topics in *Del.icio.us*. Here we present top 10 words with highest probabilities in randomly selected five topics. As we can see from each topic, the words more or less represent the semantic meaning of each topical category. For example, in TOPIC5, the words *marketing*, *advertising*, *startup* and etc. are centered on the *business* topic.

We depict the histograms of the diversity scores of users for *Last.fm*, *Flickr* and *Del.icio.us* in Fig. 6, 7 and 8 respectively, where we randomly sample 10,000 users for *Last.fm* and 1,000 users for both *Flickr* and *Del.icio.us*. The diversity scores of each user are normalized to discrete integers before we draw the histograms. From the histograms, we observe the distributions of diversity scores exhibit normal distribution property. Notice that a diversity score of zero represents the user only annotated only one item, which is usually not sufficient for characterization of interest diversity. This corresponds to the cold start problem [8].

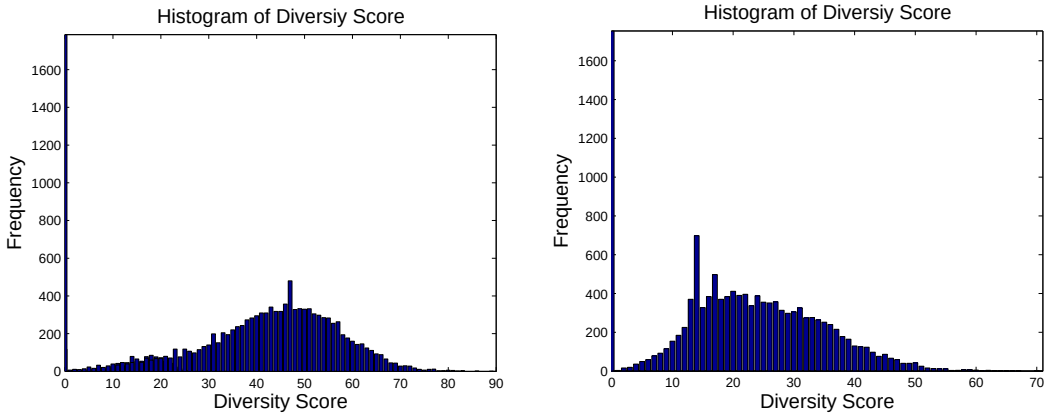
5 FRIENDSHIP PREDICTION

In this subsection, we investigate diversity features and the interplay of associated users with various degrees of interest diversity in friendship formation. Here we randomly samples 10,000 friend pairs and 10,000 non-friend pairs in *Last.fm* for experiments. Firstly, we analyze the predictive power of diversity features as well as common features. For each user pair (A, B), *diversity features* include the diversity scores of A and B, as well as the distance of the variance vectors associated to the topic representations of each item annotated by A and B. We also use the distance of the mean vectors associated to the topic representations of each item annotated by A and B. For *common features*, we use the number of shared friends, the number of shared items as well as the number of shared distinct tags between A and B.

In experiments, we adopt leave-one-out cross validation, which is to predict one missing link using all the other link information in the network [10]. A linear regression model [2] is trained on the above-mentioned features. Table 4 shows the prediction accuracy for different features. As we can see, the prediction power of diversity features are comparable with the common features as we set a relative larger number of topics ($T = 100$) in LDA. The combination of diversity and common features achieves even better performance. This demonstrates the robustness of diversity features. And it also reflects that the characterization of diverseness of individual tastes can indeed help friendship prediction.

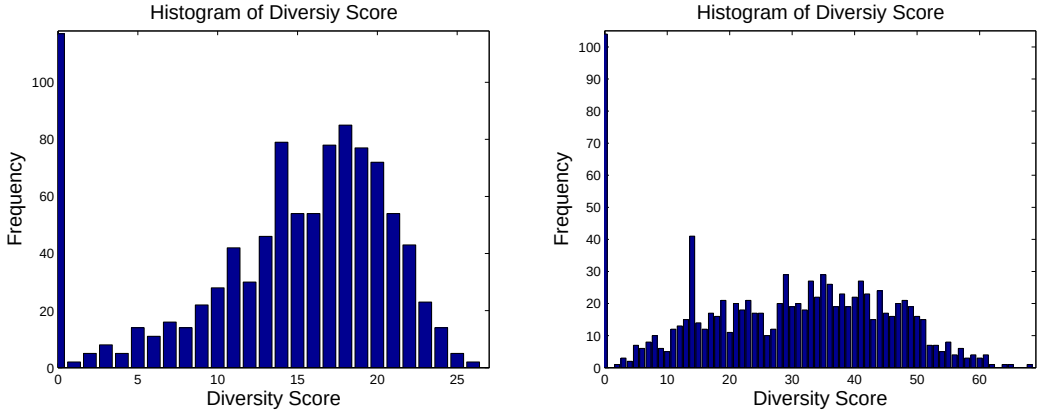
We then investigate the effect of different diversity levels of individual interests on friendship formation. Specific questions to answer include when *diversity* (user with a diversity of tastes) meets *speciality* (user with a strong propensity toward special genres of items), how easily they will become friends. We first denote each user based on his/her diversity score as follows:

ISBN: 978-1-62561-004-1



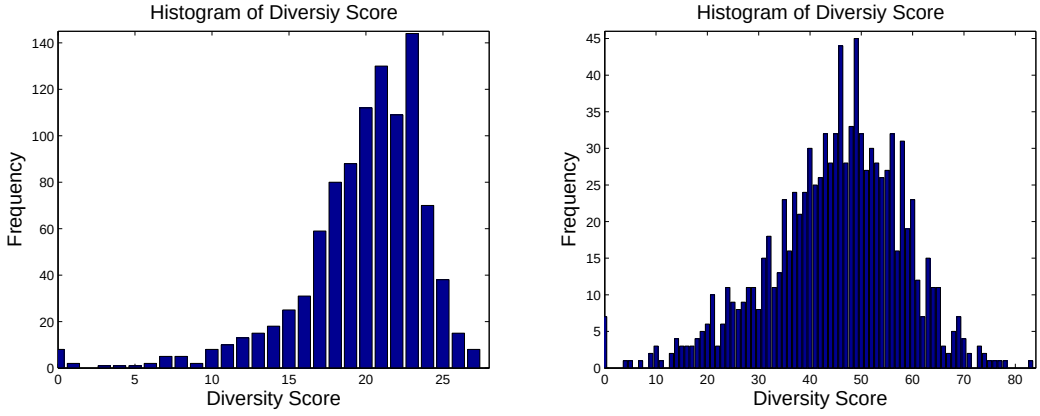
(a) $T = 10$ (b) $T = 100$

Figure 6: Distribution of diversity scores for *Last.fm*, which are generated using LDA with 10 topics and 100 topics)



(a) $T = 10$ (b) $T = 100$

Figure 7: Distribution of diversity scores for *Flickr*, which are generated using LDA with 10 topics and 100 topics



(a) $T = 10$ (b) $T = 100$

Figure 8: Distribution of diversity scores for *Del.icio.us*, which are generated using LDA with 10 topics and 100 topics

ISBN: 978-1-62561-004-1

Table 4: Accuracy of friendship prediction for *Last.fm*

# of topics	Diversity features	Common features	Diversity + Common features
10	64.4%	70.3%	70.5%
100	66.3%	70.3%	71.5%

Table 5: Accuracy of friendship prediction for users with different diverseness levels of individual interests in *Last.fm*

# of topics	(L, L)	(L, M)	(L, H)	(M, M)	(M, H)	(H, H)
10	72.8%	69.9%	72.2%	70.1%	72.3%	72.7%
100	72.8%	73.1%	70.0%	72.1%	71.8%	76.7%

- **L**: Low-level interest diversity (score ≈ 0)
- **M**: Median-level interest diversity (sore $\leq$ median score)
- **H**: High-level interest diversity (sore $\geq$ median score)

The friendships in *Last.fm*, *Flickr* and *Del.icio.us* are mutual, which means if A is a friend of B, then B is a friend of A. Hence, there are six different combinations of user pairs (A, B) for different diversity levels of individual tastes. We train individual regression model for each combination and conduct the friendship prediction respectively. Table 5 shows the results. We can observe that when users with the same level of interest diversity meet, especially for (L, L) and (H, H), the prediction yields relative high accuracy. Whereas the accuracy drops when users with different levels of interest diversity come together. The friendship prediction on (H, H) achieves the highest accuracy. It is consistent with the intuition that users with high-degree diversity of individual tastes have more tolerance on friends making, especially when they meet the same type of persons who also share a diversity of interests.

III CONCLUSION AND FUTURE WORK

We have explored the engagement of user annotation, the eccentricity of user's making friends and the diversity of individual tastes in online social networks. The investigation has sociological and psychological implications such as how the interplay of users with different diversity degrees accounts for the formation of social links and interest communities, as well as how the observations can help friend recommendation in both online social networks and in reality. For future work, we plan to investigate how the gender difference accounts for different patterns of friendship formation. Another interesting direction is to investigate the effect of role difference in diffusion of information for friendship formation, where we can consider each user as leader, early adopter or late adopter of information.

IV ACKNOWLEDGEMENT

This work was supported in part by the US National Science Foundation under grant CCF-1048311. We would like to thank Yan Liu for helpful discussions and insightful suggestions.

References

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006.

[2] C.M. Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

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

[4] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 201–210. ACM, 2009.

[5] G. Drury. Opinion piece: Social media: Should marketers engage and how can it be done effectively? *Journal of Direct, Data and Digital Marketing Practice*, 9(3):274–277, 2008.

[6] S. Goel, A. Broder, E. Gabrilovich, and B. Pang. Anatomy of the long tail: ordinary people with extraordinary tastes. In *Proceedings of the third ACM*

ISBN: 978-1-62561-004-1

international conference on Web search and data mining, pages 201–210. ACM, 2010.

[7] T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228, 2004.

[8] V. Leroy, B.B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 393–402. ACM, 2010.

[9] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. *Knowledge Discovery in Databases: PKDD 2005*, pages 133–145, 2005.

[10] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.

[11] X. Li, L. Guo, and Y.E. Zhao. Tag-based social interest discovery. In *Proceedings of the 17th international conference on World Wide Web*, pages 675–684. ACM, 2008.

[12] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. In *Conference on Information and Knowledge Management (CIKM¡¯03)*, pages 556–559. Citeseer, 2003.

[13] E. Santos-Neto, D. Condon, N. Andrade, A. Iamnitchi, and M. Ripeanu. Individual and social behavior in tagging systems. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 183–192. ACM, 2009.

[14] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies: social link prediction from shared metadata. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 271–280. ACM, 2010.

[15] X. Shi, J. Zhu, R. Cai, and L. Zhang. User grouping behavior in online forums. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 777–786. ACM, 2009.

[16] M. Slaney and W. White. Measuring playlist diversity for recommendation systems. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 77–82. ACM, 2006.

[17] B. Taskar, M.F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *NIPS*, 2003.

[18] S.H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 537–546. ACM, 2011.

ISBN: 978-1-62561-004-1