CSD: a multi-user similarity metric for community recommendation in online social networks

Xiao Han, Leye Wang, Reza Farahbakhsh, Angel Cuevas Rumin, Rubén Cuevas Rumin, Noel Crespi, Lina He

Han, X., Wang, L., Farahbakhsh, R., Cuevas, N., Cuevas, R., Crespi, N. & He, L. (2016). CSD: A multi-user similarity metric for community recommendation in online social networks. *Expert Systems with Applications, 53*, 14–26.

DOI: 10.1016/j.eswa.2016.01.003

CSD: A Multi-User Similarity Metric for Community Recommendation in Online Social Networks

Xiao Han[a,b,*], Leye Wang[b], Reza Farahbakhsh[b], Ángel Cuevas[b,c], Rubén Cuevas[c], Noel Crespi[b], Lina He[d,e]

[a] *Shanghai University of Finance and Economics, Shanghai 200433, China*
[b] *Institut-Mines Télécom, Télécom SudParis, 9 rue Charles Fourier, 91011 Evry Cedex, France*
[c] *Universidad Carlos III de Madrid, Av de la Universidad, 30 28911 Leganés, Madrid, Spain*
[d] *School of Earth Science and Engineering, Hohai University, Nanjing 210098, China*
[e] *State Key Laboratory of Geo-information Engineering, Xi'an 710054, China*

Abstract

Communities are basic components in networks. As a promising social application, community recommendation selects a few items (e.g., movies and books) to recommend to a group of users. It usually achieves higher recommendation precision if the users share more interests; whereas, in plenty of communities (e.g., families, work groups), the users often share few. With billions of communities in online social networks, quickly selecting the communities where the members are similar in interests is a prerequisite for community recommendation. To this end, we propose an easy-to-compute metric, *Community Similarity Degree* (CSD), to estimate the degree of interest similarity among multiple users in a community. Based on 3460 emulated Facebook communities, we conduct extensive empirical studies to reveal the characteristics of CSD and validate the effectiveness of CSD. In particular, we demonstrate that selecting communities with larger CSD can achieve higher recommendation precision. In addition, we verify the computation efficiency of CSD: it costs less than 1 hour to calculate CSD for over 1 million of communities. Finally, we draw insights about feasible extensions to the definition of CSD, and point out the practical uses of CSD in a variety of applications other than community recommendation.

Keywords: Online Social Network, Community Similarity Degree, Community Recommendation, Community Selection

1. Introduction

With the overwhelming explosion of Online Social Networks (OSNs), a large number of online communities are naturally formed by people who share certain properties. As reported, Google was able to index 620 million user-created communities in Facebook by 2010[1]; Orkut exhibits more

*Corresponding author. Tel.:+86 21 6590 1498
Email addresses: xiaohan@mail.shufe.edu.cn (Xiao Han), leye.wang@telecom-sudparis.eu (Leye Wang), reza.farahbakhsh@telecom-sudparis.eu (Reza Farahbakhsh), acrumin@it.uc3m.es (Ángel Cuevas), rcuevas@it.uc3m.es (Rubén Cuevas), noel.crespi@telecom-sudparis.eu (Noel Crespi), hlnyh@hhu.edu.cn (Lina He)

[1] http://allfacebook.com/google-now-indexes-620-million-facebook-groups_b10520

than 100 million communities along with hundreds of newly created communities every day (Chen et al., 2008). This huge number of online communities and the common properties of users within communities have led to a new paradigm of recommendation systems through OSNs, namely community recommendation.

Community recommendation suggests particular items (e.g., movies, music, books) to a group of users and aims to convince the users to adopt its recommended items; it will achieve better performance if more users are interested in the recommended items (i.e., higher recommendation precision) (Gorla et al., 2013; Hu et al., 2014). Instead of targeting an individual user, community recommendation presents many advantages. First, as human beings are of a social nature, recommendation for users within a community is required in some cases (Hu et al., 2014), such as recommending a tourist attraction to a group of friends to spend holiday, or advertising to community forums in OSNs. Second, community recommendation may also be conductive to address *new-user* problem in recommendation systems by recommending the new users items based on the interests of other users in the same community (Masthoff, 2011). Moreover, since recommending items to a community merely requires the community's collective interest information but not necessarily every user's personal interests (Aimeur et al., 2006), community recommendation can preserve privacy for the users who are unwilling to reveal personal information by certain approaches such as obfuscating interests of users in a community (Parameswaran & Blough, 2007).

Concerning the potential benefits of community recommendation, much existing work puts effort to devise sophisticated algorithms for selecting items that are probably preferred by most users in a given community (Baltrunas et al., 2010; Gorla et al., 2013; Hu et al., 2014). However, recall that there exist millions of communities with various natures in OSNs, whether a sophisticated algorithm can recommend satisfactory items to all the communities is in doubt. Intuitively, some communities in which users share many interests may be intrinsically appropriate for community recommendation to achieve high recommendation performance; while for some other communities consisting of users with distinct interests (e.g., a Random-based community of people for a statistic survey), a community recommendation system with sophisticated algorithms may still hardly find any items that are preferred by most users in such a community. In order to avoid unduly running sophisticated recommendation algorithms for the inappropriate communities, in this paper, we investigate how to quickly select the appropriate communities in which community recommendation may achieve high performance, from millions of communities in OSNs.

To address this issue, we rely on the principle that a community is more effective for recommendation if the members in the community present more common interests (Baltrunas et al., 2010); hence, we propose to measure the interest similarity among users in a community and then select the communities of a large similarity degree as the appropriate ones for community recommendation.

Although the basic idea seems straightforward, it is non-trivial to be implemented. First, we need to measure interest similarity among multiple users in a community. Many similarity measurements (Spertus et al., 2005) have been proposed; whereas most of them focus on the similarity between two individuals rather than among multiple users. Second, the interest similarity measure should be efficient to compute, so that it can be fast enough to select the appropriate communities over a huge number of ones in real-life OSNs.

For our purposes, firstly, we define a metric — *Community Similarity Degree* (CSD) — to compute the degree of similarity among the users in a community based on their common interests. The CSD value ranges from 0 when the users in a community do not share any interest, to 1 if all

the users present exactly the same interests.

Subsequently, with $208K$ user profiles collected from Facebook, we conduct extensive empirical studies to understand the properties of CSD by emulating four types of communities (i.e., *Friend-based*, *Interest-based*, *Location-based* and *Random-based communities*). We observe that CSD decreases with the increase of either the number of users or the number of interests. We also notice that the Interest-based communities which are formed by users having one common interest normally exhibit $1.45\times$ to $4.5\times$ larger CSD than the Friend- or Location-based communities where users share one friend or come from the same city. As we exclude the common interest in an Interest-based community to calculate its CSD, this observation indicates that users with one common interest are likely to share more other interests than friends or people in the same city.

Finally, with a simulated community recommendation system, we validate the effectiveness and efficiency of CSD in community selection. We demonstrate that selecting the communities with large CSD can achieve good recommendation performance, i.e., high average recommendation precision. We also compare different average precisions when the recommendation is respectively applied to Interest-, Friend-, Location-, and Random-based communities. The experiment results confirm that the Interest-based communities, which have larger CSD, gain $2\times$ higher median average precision when it compares to Friend-, Location-, Random-based communities. This result further indicates that selecting communities with large CSD is effective to achieve good performance in community recommendation. Moreover, we verify the computation efficiency of CSD and demonstrate that we can compute CSD for 1 million of communities within 41 minutes.

In summary, the main contributions of this paper are: (i) We define a metric called CSD to estimate interest similarity degree among multiple users within a community, while most of the existing similarity metrics compute the similarity between two objects. (ii) We conduct extensive empirical studies on a large real Facebook dataset and reveal CSD's characteristics based on 3460 emulated communities. (iii) We emulate a community recommendation system and demonstrate that CSD is an effective and efficient metric to select the appropriate communities for community recommendation. (iv) We give insights about feasible extensions to the definition of CSD and present practical uses of CSD in various applications besides community recommendation.

The rest of this paper is organized as follows. Section 2 reviews some related work. Section 3 defines the metric of CSD. Section 4 introduces our dataset and the communities constructed based on the dataset. We conduct empirical studies of CSD in Section 5. In Section 6, we emulate a community recommendation system and validate the effectiveness and efficiency of CSD. Finally, Section 7 concludes the paper.

2. Related Work

In this section, we briefly review the existing related work through two aspects: (i) recommendation systems; (ii) similarity metrics and the use of similarity in social applications.

2.1. Recommendation Systems

Recommendation systems are extremely promising for marketing in OSNs by providing users with suggestions, such as what products to purchase, what movies to watch or what books to read (Ricci et al., 2011). Much work proposes various approaches (e.g., hierarchical Bayesian model (Purushotham et al., 2012), trust circle-based model (Yang et al., 2012), semantic similarity-based model (Dong et al., 2011)) to provide personalized recommendations to users. Such recommendation systems are normally classified into three categories according to the ways of recommen-

dation, including content-based, collaborative and hybrid recommendation approaches (Adomavicius & Tuzhilin, 2005). Most of these systems concentrate on recommendation for an individual user (Deng et al., 2014); however, our work tends to improve community recommendation which recommends items for a group of users instead of an individual.

In recent years, some studies have proposed to select items for a community of users. Baltrunas et al. (2010) exploit a collaborative filtering algorithm to generate personalized recommendations for an individual user and then leverage a rank aggregation method to produce a joint ranking list of recommendations for a community of users. Focusing on better modeling the users within a community, Gorla et al. (2013) design a probabilistic community recommendation method to improve the aggregation of individuals' recommendations. By considering the collective features that may determine users' choices within a community, Hu et al. (2014) propose a joint community recommendation model which accommodates both users' individual interests and community decision. Various methods, including content-based, user-based and hybrid of content and user, for producing recommendations for a community of users are examined and compared by Ronen et al. (2014). These sophisticated algorithms and models concentrate on how to select the items for a given community of users, whereas it may work inefficiently if the users in a community do not share many interests. To tackle this issue, in this paper, instead of designing a recommendation algorithm, we attempt to find the communities that can achieve good performance in community recommendation.

With a similar research objective as our work, recently Basu Roy et al. (2015) study how to form communities so that most users in the formed communities are satisfied with the recommendations; while the difference between our work and Basu Roy et al. (2015) is still significant: rather than designing community formation algorithms to create new communities, we define an effective and efficient metric, CSD, to select the appropriate communities from a huge number of self-organized communities that have already existed in real-life networks nowadays.

2.2. Similarity Metrics in OSNs

Evaluating similarity is a practical and fundamental problem with a long history, which serves in various research domains such as geographic information science (Schwering, 2008), biology (Lei et al., 2013), and decision-making (Tsebelis, 1995). In OSNs, a series of classical metrics, including overlap, cosine similarity, Jaccard similarity, Pearson correlation coefficient, etc., are employed to estimate the strength of user relationships, the similarity of users' tastes/interests, and the resemblance of users' background (Han et al., 2014, 2015; Sarwar et al., 2001). To recommend social events with holding a user's home location, the location similarity is calculated by weighted cosine similarity taking into account the common events that users from both locations have attended (Quercia et al., 2010). Besides, Han et al. (2014) study similarity between two users by both common friends and common interests and show that friends generally share more interests than strangers. Pearson correlation coefficient is rather popular in collaborative filtering recommendation systems as it subtracts the average rating score from each rating, thereby eliminates the individual subjective differences (Sarwar et al., 2001).

Semantic objects, such as comments, posts, answers to questions, descriptions or reviews about services/products, and tags to photos, videos, music, are widespread over OSNs nowadays. Estimating two users' similarity by their semantic relatedness is a fundamental task, which can in turn support a great number of applications (e.g., recommendation system, information retrieval, and link prediction) (Markines & Menczer, 2009). Accordingly, similarity metrics, such as mutual information (Hindle, 1990), Lin's descriptive similarity (Lin, 1998), and maximum information

path (Markines & Menczer, 2009), are proposed to capture the structural information between semantic objects.

Besides, a collection of global structural similarity metrics (e.g., Katz, PageRank) are proposed to capture the global topology information based on structural network. These metrics are widely-used to measure the similarity in link prediction, trust estimation, and community detection. To predict the structure of social network without knowing any author-author relationships, Makrehchi (2011) constructs auxiliary networks based on author-topic and topic-topic relations and uses Katz metric to calculate the closeness of either author-topic or topic-topic relations. Backstrom & Leskovec (2011) calculate PageRank score to predict and recommend links in a supervised way. Rossi et al. (2015) survey the existing graph-based and feature-based similarity methods for role discovery in networks, and propose a flexible framework for discovering roles using the notion of similarity on a feature-based representation.

Recently, with the arrival of the big data era, a real-life network can grow up to billions of nodes and edges. Thus, improving the computation efficiency and scalability of similarity metrics begins to attract much research interest. Kusumoto et al. (2014) propose a fast and scalable algorithm to compute the top-k similar nodes for a given node in terms of the SimRank metric; while Tao et al. (2014) design an efficient algorithm to select the k most similar pairs of nodes with the largest SimRank similarities among all possible pairs. Zhang et al. (2015) use the idea of random path to quickly select the top-k similar nodes for a given node in a huge network and applies this method in two applications — identity resolution and structural hole spanner finding. In our definition of CSD, we also consider the computation efficiency so that we can use CSD for selecting appropriate communities for recommendation from millions of communities in a reasonable time period.

In summary, most of these state-of-the-art works focus on the metrics considering the similarity between two users, whereas this paper intends to compute the similarity among a community of users.

3. Community Similarity Degree

In this section, we define Community Similarity Degree (CSD) to measure the interest similarity among users in a community. We start with some intuitive concepts about interest similarity of a set of users (or called community members). Then, we introduce some assumptions and criteria to formulate the similarity intuitions. Finally, we give the definition of the metric (CSD) to meet the established criteria based on the assumption. During the metric definition, we note that: (1) if a user reports an interest we call the user a *fan* of the reported interest; (2) we aggregate all the users' interests and construct an interest set for a community. Each element in the interest set is a distinct interest.

Before defining the metric, we first clarify some intuitions of similarity among multiple users. The being defined similarity metric is expected to capture the following intuitions.

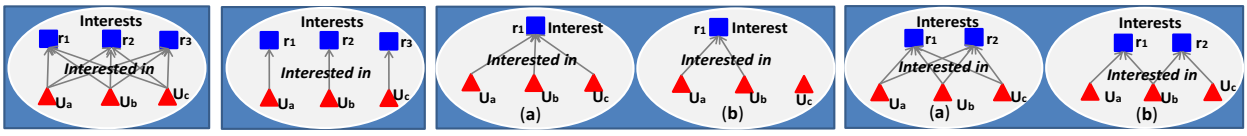
- **Intuition 1:** If all the community members exhibit exactly the same interests, their interest similarity reaches the highest value.
- **Intuition 2:** If any two members share no interest, the interest similarity of the community should be the lowest value.
- **Intuition 3:** Assume only one distinct interest is reported in a community with a certain number of users, then the more fans the distinct interest has (some users may report no interest), the higher the interest similarity is.

- **Intuition 4:** Given a community with a certain number of members and distinct interests, the interest similarity of the community should be higher if there exist more fans for every single distinct interest (i.e., the sum of the fan number for each distinct interest is larger).

To formulate the intuitions and define the metric, we introduce some notations here. We mark a given community as $c = \{\mathcal{U}_c, \mathcal{R}_c\}$, where $\mathcal{U}_c$ represents the users in the community and $\mathcal{R}_c$ stands for the set of all the users' interests (i.e., the distinct interest set in the community c). The number of users and the number of distinct interests in the community are respectively denoted as $\mathcal{N}_u(c)$ and $\mathcal{N}_r(c)$. For each distinct interest $r \in \mathcal{R}_c$, we count the number of its fans as its popularity, denoted as $p(r)$; then, we can sum the number of fans for all the distinct interests as the weight of the community, i.e., $W(c) = \sum_{r \in \mathcal{R}_c} p(r)$.

Following the intuitions, we establish the following assumptions and criteria:

- **Assumption:** We assume the highest value of the being defined metric CSD is 1 while the lowest value equals 0.
- **Criterion 1:** When all the users in a community have exactly the same interests, i.e., all the users are interested in each distinct interest in $\mathcal{R}_c$, CSD is 1; i.e., $\text{CSD}(c) = 1$, *iff:* $W(c) = \sum_{r \in \mathcal{R}_c} p(r) = \sum_{r \in \mathcal{R}_c} \mathcal{N}_u(c) = \mathcal{N}_r(c) \times \mathcal{N}_u(c)$. (According to **Intuition 1**)
- **Criterion 2:** When any two members do not share any interest, i.e., each distinct interest only has one fan, CSD is 0; i.e., $\text{CSD}(c) = 0$, *iff:* $W(c) = \sum_{r \in \mathcal{R}_c} p(r) = \sum_{r \in \mathcal{R}_c} 1 = \mathcal{N}_r(c)$. (According to **Intuition 2**)
- **Criterion 3:** Given two communities c_1 and c_2 with the same number of users (i.e., $\mathcal{N}_u(c_1) = \mathcal{N}_u(c_2) = \mathcal{N}_u$) and the same number of distinct interests (i.e., $\mathcal{N}_r(c_1) = \mathcal{N}_r(c_2) = \mathcal{N}_r$), then the community presenting the larger weight has the larger CSD; i.e., $1 \geq \text{CSD}(c_1) > \text{CSD}(c_2) \geq 0$, *iff:* $\mathcal{N}_r \times \mathcal{N}_u \geq W(c_1) > W(c_2) \geq \mathcal{N}_r$. (According to **Intuition 3 and 4**)



Intuition 1: users in the community share the **most** interests (i.e., the highest CSD)

Intuition 2: users in the community share the **least** interests (i.e., the lowest CSD)

Intuition 3: users in community (**a**) share more interests than users in community (**b**), since more users in (**a**) prefers the certain interest than the users in (**b**) (i.e., CSD (**a**) > CSD (**b**))

Intuition 4: extending intuition 3 to the cases of multiple interests, users in community (**a**) share more interests than users in community (**b**) (i.e., CSD (**a**) > CSD (**b**))

Figure 1: Examples of Intuitions and Criteria for CSD Definition.

An example is provided in Figure 1 to better illustrate the intuitions and criteria. Suppose that we have a community with three users u_a, u_b and u_c. If all these users report that they are interested in r_1, r_2 and r_3, then their interests are exactly the same and the interest similarity of this community should be the highest (**Intuition 1/Criterion 1**). Whereas, if u_a prefers r_1, u_b likes r_2 and u_c favors r_3, then there is no common interest among u_a, u_b and u_c. In this case, the interest similarity of the community is the lowest (**Intuition 2/Criterion 2**). Assume r_1 is the only reported interest by the community members. The community interest similarity of the case

that three users all prefer r_1 is higher than another case that u_a and u_b are interested in r_1 but u_c does not claim any preference (**Intuition 3/Criterion 3**). On the basis of Intuition 3 where only one interest is considered, we can extend the case to multiple interests. Specifically, compared to a community where all the three users u_a, u_b and u_c like interests r_1 and r_2, another community in which u_a and u_c prefer r_1 meanwhile u_b and u_c favor r_2 presents lower interest similarity (**Intuition 4/Criterion 3**).

Definition: Based on the assumption, while meeting the aforementioned criteria, we define CSD as:

$$\mathrm{CSD}(c) = \frac{W(c) - \mathcal{N}_r(c)}{\mathcal{N}_u(c) \times \mathcal{N}_r(c) - \mathcal{N}_r(c)} = \frac{W(c)/\mathcal{N}_r(c) - 1}{\mathcal{N}_u(c) - 1} \tag{1}$$

where $W(c)/\mathcal{N}_r(c)$ calculates the average popularity of interests in community c. Therefore, in other words, CSD assesses the interest similarity of users in a community approximately by the ratio between the average popularity of interests ($W(c)/\mathcal{N}_r(c)$) and the total number of community members ($\mathcal{N}_u(c)$). We note that the value of CSD ranges from 0 to 1.

In addition, CSD is an easy-to-compute metric. The computation complexity of CSD is $O(N)$, where N denotes the number of users in a community, because we only need to enumerate all the users' interests once to calculate CSD. Comparatively, if we estimate the community interest similarity by computing the conventional pairwise interest similarity (e.g., cosine similarity) between any two users and then averaging all the pairwise similarities, the complexity would be $O(N^2)$ as the total number of user pairs in a community is $N(N-1)/2$.

4. Data and Community Description

In this section, we briefly introduce our dataset and its collection procedure. We also describe four different types of communities created with our collected dataset.

4.1. Data Description

In order to validate our proposed metric, we have developed a web crawler to collect users' information from Facebook. Given one root user, the crawler then follows the Breadth-First Search (BFS) approach (Gjoka et al., 2011) to go through the user's friends and friends of friends (i.e., two-hop friends). For each user/friend, the crawler captures the user profile which includes the user's demographic information (e.g., birthday, gender, home town) and interests in terms of five well-defined categories (i.e., television, books, music, movies and games) (Han et al., 2015). Note that we respect the users' privacy by collecting only their public information and anonymizing the user IDs.

To count, the collected Facebook dataset contains $208,634$ users and $542,597$ distinct interests from the above-mentioned five categories. In our dataset, the users present 11 interests on average; and 12% of the users only report one single interest, while 5% of them include more than 100 interests. Furthermore, the users in our dataset are from more than 150 countries and $9K$ cities; meanwhile there exists a wide variability in the number of users' friends, which ranges from dozens to $5K$. Finally, we note that most of the collected distinct interests show a small popularity (85% of the distinct interests have fewer than 10 fans). However, we still find more than $10K$ and $1K$ distinct interests with more than 100 and $1K$ fans respectively; the top 100 interests are shared by more than $8K$ users.

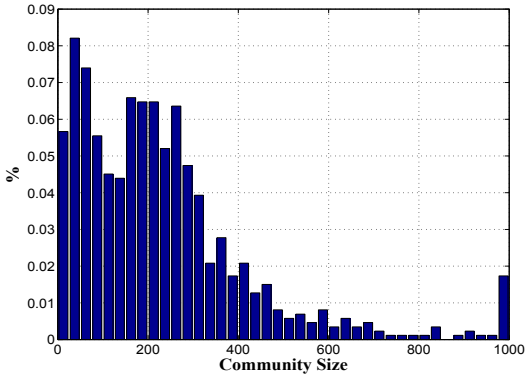


Figure 2: Community size distribution for the four analyzed community types.

4.2. Community Description

In order to investigate the characteristics of CSD and evaluate its effectiveness and efficiency in community selection for recommendation, we construct four different types of communities using our collected Facebook dataset: Friend-based, Random-based, Interest-based and Location-based communities. Next we describe how each type of community is generated.

4.2.1. Friend-based Communities

A Friend-based community is formed by a user and all her friends with at least one interest. Creating these communities was straightforward due to the BFS technique employed by our crawler. Our dataset allowed us to define 865 Friend-based communities. It is worth mentioning that Friend-based communities are typically used for recommendation purposes in OSNs (Chen et al., 2009). For instance, when a Facebook user starts utilizing an application, her friends may receive a notification indicating that fact. This notification/recommendation is based on the belief that two friends likely have similar tastes.

Finally, as we will see later, the community size (i.e., the number of users) has a direct impact on the CSD value for the communities under study. Therefore, in order to perform a fair comparison, the remaining community types replicate the community size distribution of Friend-based communities. Figure 2 shows the distribution for the size of the 865 Friend-based communities (that is the same for the other community types). Note that the rightmost bin in the figure represents all the communities whose size is larger than 1000.

4.2.2. Random-based Communities

We create 865 Random-based communities. As mentioned before, we decide the size of each Random-based community according to the size distribution of Friend-based communities. For a Random-based community of size N, we select N random users from our dataset to create the community.

4.2.3. Interest-based Communities

An Interest-based community is formed by a set of users who all present one common interest (e.g., users who are interested in a same movie). Note that each user will typically have some other interests in addition to the common one. Similarly, we generate 865 Interest-based communities by following the same size distribution of Friend-based communities. Given a size of N users to

generate an Interest-based community, we find the list of all those interests whose popularity is N (i.e., N users present that interest in their profile) and randomly select one interest to construct the community with all its fans.

4.2.4. Location-based Communities

A Location-based community is formed by all the users showing the same *Current City* attribute. As Facebook allows users freely to input any text in their profile attributes, a city can be marked by several diverse notations (e.g., New York, New York City, NY, NYC, etc. all indicate the same city). Therefore, we use *Yahoo PlaceFinder API*[2] to unify all the different notations for a city and obtain 9K unique cities. Accordingly, the users in a same unique city are grouped into a community. We then select 865 Location-based communities with the same size distribution as the other three types of communities.

5. Empirical studies on CSD

In this section, we carry out extensive empirical studies on CSD. We first study how CSD varies with number of users ($N_u(c)$), number of interests ($N_r(c)$) and community weight ($W(c)$) respectively. Then we compare the distributions of CSD by the four community types. We further look into CSD of Interest-based communities by different interest categories in the end.

5.1. CSD characterization

As we have seen in Section 3, CSD is a metric varying with the community weight, the number of users (i.e., community size) and the number of interests. Here we study how each of these factors would influence CSD. We conduct the investigations separately on four types of communities and obtain the similar conclusions. For the sake of brevity, we only show the results of Friend-based communities as a representative.

5.1.1. Influence of Number of Users

This subsection analyzes the impact of the number of users on CSD. We group communities of similar sizes into bins. In particular, we use two methods to construct the bins, *equal width binning* and *equal frequency binning* (Dougherty et al., 1995).

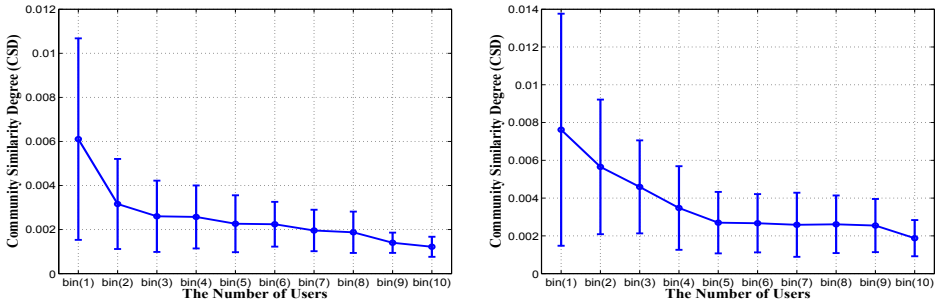
In equal width binning, we consider 10 different bins such that $bin(b)$ includes all those communities whose size belongs to the interval $((b-1)*100, b*100]$, for b going from 1 to 9. Hence, the first bin ($b = 1$) includes communities with size in the interval $[2, 100]$[3], the second bin is $[101, 200]$, and so on. The last bin ($bin(10)$) includes communities with size larger than 900 users.

In equal frequency binning, we also generate 10 bins where each bin contains approximately 10% of all the communities. Specifically, we first rank all the communities according to the number of users ascendingly. Then the first bin includes the first 10% of communities in the sorted list, the second bin includes 10-20%, and so on. Note that the number of users of the communities in $bin(b)$ is large than the ones in $bin(b-1)$.

Figure 3(a) presents the average CSD and the corresponding standard deviation by the number of users of communities with equal width binning. The results show that CSD decreases with the

[2] https://developer.yahoo.com/boss/placefinder/

[3] It does not make sense to evaluate communities with a single user.



(a) Equal width binning

(b) Equal frequency binning

Figure 3: CSD vs. community size. In 3(a), the size of communities in $bin(b)$ ($b \in [2, 9]$) belongs to $((b-1)*100, b*100]$; Particularly, the communities in $bin(1)$ have a size belonging to $[2, 100]$ while the size of communities in $bin(10)$ is larger than 900; In 3(b), each bin contains the same number of communities and the size of communities in $bin(b)$ is larger than $bin(b-1)$.

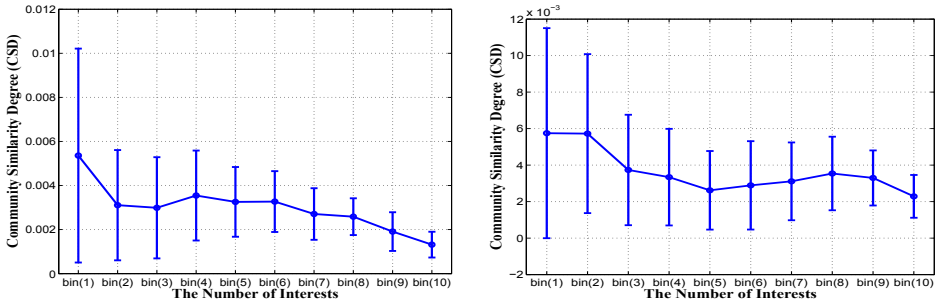
increase of the community size. The major drop appears between the two first bins where the average CSD in $bin(1)$ doubles $bin(2)$; while the drop gets stable as the community size increases. Figure 3(b), with equal frequency binning, shows the similar trend. This is an expected result since larger communities present more users (by definition), and then generally contain a larger number of interests. In particular, for the 865 Friend-based communities, we have found (by using a linear regression model) that the number of available interests in a community is roughly $12\times$ the size of the community. Therefore, bigger communities typically bring a larger diversity of both users and interests, which intuitively leads to a lower similarity degree (CSD). Although there is an obvious drop trend, a community with a larger size does not necessarily achieve a lower CSD and small communities also probably have a very low CSD. This is demonstrated by the large standard deviation of CSD, especially for the bins of smaller sizes (e.g., $bin(1)$).

5.1.2. Influence of Number of Interests

We now analyze how the number of interests in the community impacts CSD. We repeat the methodology used in the previous subsection and group communities with similar number of interests in 10 different bins. With equal width binning, $bin(b)$ includes those communities with a number of interests within the interval $((b-1)*1000, b*1000]$ for values of b ranging from 1 to 9, while the last bin ($b = 10$) includes those communities with more than 9000 interests. With equal frequency binning, the $bin(1)$ includes the first 10% of communities in the ranking list sorted ascendingly by the number of interests, the $bin(2)$ includes 10-20%, and so on.

Figure 4(a) plots the average CSD with its standard deviation by the number of interests with equal width binning. Not like the steady drop of CSD by community size, the figure presents a decreasing trend of CSD with micro-fluctuations by the number of interests. For instance, it shows a slight but obvious increase of average CSD of communities when the number of interests in communities grows from $(2K, 3K]$ to $(3K, 4K]$. The change of standard deviation in the figure indicates that the variability of CSD gets smaller with the increase of the number of interests. According to Figure 4(b) with equal frequency binning, we can also find a similar non-steady drop trend of CSD as the number of interests increases.

The observed behavior can be explained since, in the case of the interests there are two important aspects to consider. On one hand, a higher number of interests lead to a higher diversity,



(a) Equal width binning (b) Equal frequency binning

Figure 4: CSD vs. the number of interests. In 4(a), the number of interests of communities in $bin(b)$ ($b \in [1, 9]$) belongs to $((b-1)*1000, b*1000]$; while the communities in $bin(10)$ have interests larger than 9000; In 4(b), each bin contains the same number of communities and the number of interests of communities in $bin(b)$ is larger than $bin(b-1)$.

which tends to reduce the CSD. This factor is responsible for the general reduction trend. On the other hand, with the increase of the number of interests, the total popularity of all the interests within the community (i.e., community weight) would probably increase as well, which may produce the flat or increasing evolution of CSD between some bins[4]. We further explore the factor of community weight in the next subsection.

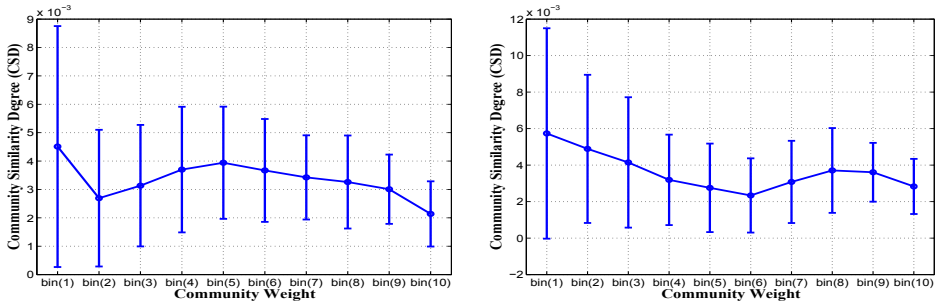
5.1.3. Influence of Community Weight

In this subsection, we discuss the impact that the community weight has on the CSD. We used the same technique as in the previous subsections. With equal width binning, we have 10 bins so that $bin(b)$ includes those communities having a weight within the interval $((b-1)*2000, b*2000]$ with b ranging from 1 to 9. The last bin ($b = 10$) includes all the communities with a weight larger than 18000. With equal frequency binning, we sort all the communities ascendingly by their community weights and then group them into 10 bins, each of which includes approximately 10% of all the communities.

It is worth noting that, in the definition of CSD, community weight ($W(c)$) plays a different role from the number of users ($\mathcal{N}_u(c)$) and the number of interests ($\mathcal{N}_r(c)$). Recall that CSD is defined as: $\frac{W(c)/\mathcal{N}_r(c)-1}{\mathcal{N}_u(c)-1}$. Thus, by definition, $W(c)$ is positively correlated with CSD, while both $\mathcal{N}_u(c)$ and $\mathcal{N}_r(c)$ are negatively correlated with CSD. This means that the increase of $W(c)$ may indicate the increase of CSD, instead of the drop. Now let us see whether this trend can happen in real-life communities.

Actually, against the above intuition, from both Figure 5(a) and 5(b), we can hardly find the increasing trend of CSD as the community weight increases. However, we notice that Figure 5(a) and 5(b) are quite similar to Figure 4(a) and 4(b), respectively. The possible explanation for the high similarity between these figures is that community weight $W(c)$ and the number of interests $\mathcal{N}_r(c)$ are highly correlated in real-life communities. To verify whether such high correlation exists, we calculate the Pearson correlation coefficient between $W(c)$ and $\mathcal{N}_r(c)$. Based on the 865 Friend-

[4]In the definition of CSD, community weight $W(c)$ is in the numerator part; thus the increase of $W(c)$ makes CSD become larger, assume the other two factors (number of users $\mathcal{N}_u(c)$ and number of interests $\mathcal{N}_r(c)$) keep unchanged.



(a) Equal width binning (b) Equal frequency binning

Figure 5: CSD vs. community weight. In 5(a), the weight of communities in $bin(b)$ ($b \in [1,9]$) belongs to $((b-1)*2000, b*2000]$; while the communities in $bin(10)$ have a weight larger than 18000; In 5(b), each bin contains the same number of communities and the weight of communities in $bin(b)$ is larger than $bin(b-1)$.

based communities, the result coefficient is 0.938, which verifies the high correlation of $W(c)$ and $\mathcal{N}_r(c)$ actually exists in real-life communities.

To conclude, the number of users in a community is the most sensitive factor related to the change of CSD. To some extent, it meets the intuition that the users easily share interests if the community size is small and the users are tight-knit; while it is hard to find a large number of users with the same preferences. The other two factors, the number of interests $\mathcal{N}_r(c)$ and the community weight $W(c)$, are highly correlated in real-life communities but have opposite effects on CSD. Therefore, in reality, the change of CSD with the increase of $\mathcal{N}_r(c)$ is similar to that with the increase of $W(c)$; however, the change does not follow a steady trend.

5.2. CSD by Different Types of Communities

Figure 6 plots the CDF of the CSD across the 865 communities within each community type. It shows that the CSD values[5] of real-life communities are rather small, compared to the maximum possible value 1 in its definition. In addition, the observation demonstrates that Interest-based communities present the largest CSD among all the four types of communities. Particularly, the median value of CSD for Interest-based communities is 3×, 2.5× and 2× larger than Random-based, Friend-based and Location-based communities, respectively. In a word, the results indicate that the absolute degree of interest sharing among real-life community members is generally low, whereas the comparative differences of CSD between different communities are still relatively obvious. In addition, it is worth noting that in order to make a fair comparison, when computing the CSD for an Interest-based community, we have excluded the common interest, which was used to build up the community, from the distinct interest set $\mathcal{R}_c$.

In addition, we study CSD of four types of communities by the community size, shown in Figure 7. We group the communities of similar sizes using the following bins: [2,100], [101,200], [201,300], ..., [>900]. We observe that the Interest-based communities can achieve 1.45× to 4.5× CSD compared to Friend/Location-based communities, while 2.5× to 7× CSD compared to Random-based communities, with respect to different community sizes.

[5]Note that respectively 100%, 100%, 100% and 92% of the Friend-based, Random-based, Location-based and Interest-based communities have the CSD less than 0.03. The highest CSD we have found among all the analyzed communities is 0.3. For the sake of clarity, Figure 6 shows the CDF for CSD values only up to 0.03.

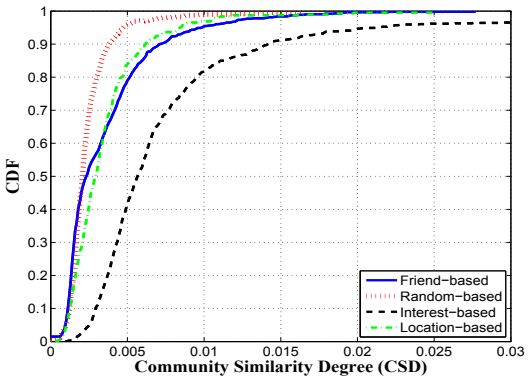


Figure 6: CDF of CSD for Friend-based, Random-based, Location-based and Interest-based communities.

In summary, these observations suggest that users with one common interest (Interest-based communities) are likely to share more other interests than friends or people from the same city (Friend/Location-based communities). Recall that our objective of proposing CSD is to select appropriate communities for recommendation, thus we expect that Interest-based communities (with higher CSD) would achieve better recommendation performance than the other types of communities. Note that this expectation will be evaluated later in Section 6.2.2.

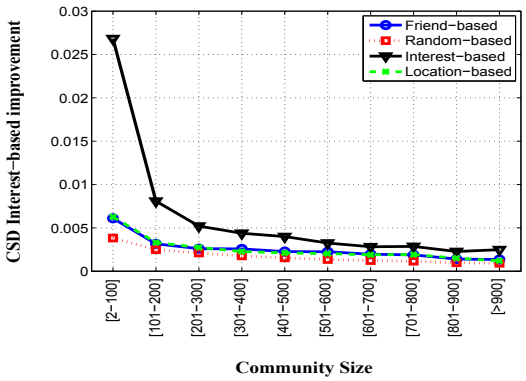
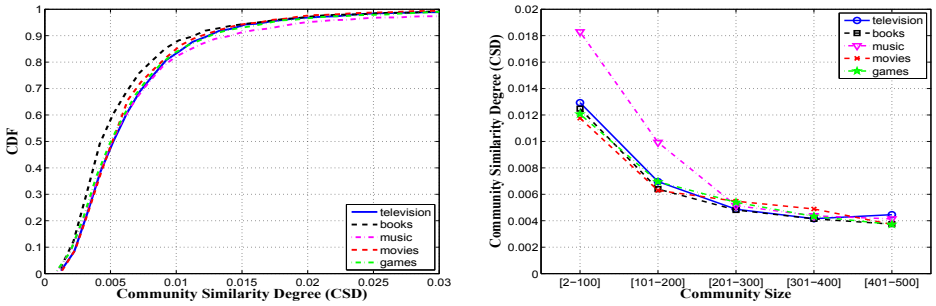


Figure 7: CSD of Interest-based, Friend-based, Location-based and Random-based communities by community sizes.

5.3. CSD by Different Interest Categories

So far, we have created Interest-based communities without paying attention to the used interest category. In this subsection, we study CSD of Interest-based communities according to different categories (i.e., television, books, music, movies, and games) of the interests that are used to create the communities. For this study, we have created 1000 communities for each category following the uniform distribution of the community size ranging from 2 to 500 users.

Figure 8(a) shows the CDF of the CSD for different interest categories. The most important observation is that all the categories follow very similar CSD distributions. The only noticeable issue is that music shows a slightly higher CSD than the remaining categories for high CSD values (which likely belong to small communities).



(a) CDF of CSD

(b) CSD by community size

Figure 8: CSD comparison among Interest-based communities by different interest categories.

Figure 8(b) shows the average CSD of different interest categories by the following community size bins: [2,100], [101,200], ..., [401,500]. The figure demonstrates that, for small communities with fewer than 200 users, music is the interest category showing the highest CSD in Facebook (and significantly higher than the other categories). The rest communities present very similar results for the first two bins. For the communities with more than 200 members, all the categories, including music, show similar behaviors.

5.4. Discussion of Observations

We observe that a majority of the emulated communities gain a very small value of CSD[6]. The small value of CSD of a community indicates that the members often share few interests. We also observe that the CSD of a community is perhaps smaller if it contains more users and interests (Figure 3 and Figure 4). Communities in OSNs usually present different characters with dozens, hundreds or thousands of users who report various interests, thus CSD of these communities may be either relatively small or large. As conducting community recommendation to the communities in which the members share fewer interests is unreasonable or costly, it is necessary to select the communities where users have relatively more common interests (i.e., with relatively larger CSD) to improve recommendation performance, especially when a huge number of communities exist. To sum up, these observations reveal the requirement of community selection approach for community recommendation in OSNs and support our original intention of proposing CSD.

6. Use Case: CSD for Community Recommendation

In this section, we validate the effectiveness and efficiency of the newly proposed CSD metric. Concerning the effectiveness, we tend to evaluate that CSD can be used to select appropriate communities which may achieve high precision in community recommendation. Specifically, we emulate a community recommendation system with our Facebook dataset and leverage an existing community recommendation approach (Baltrunas et al., 2010) to recommend items for each community; we then sort the communities by their CSD values and expect that the communities with larger CSD can achieve higher recommendation precision. For the efficiency, we will evaluate the computation time of CSD for a community.

[6]the CSD of 98% of the communities is smaller than 0.03 whereas the defined maximum value is 1

Next, we will first briefly introduce the exploited community recommendation approach and the metric for evaluation; then, we report the evaluation results.

6.1. Recommendation Approach and Metric

We implement a community recommendation system based on the idea of rank aggregation and collaborative filtering (Baltrunas et al., 2010). This approach contains two steps: first, it computes a recommendation ranking list for each individual user in the community; then, it aggregates all the users' individual recommendation ranking lists via certain pre-defined heuristics and generates an aggregated recommendation ranking list for the community.

To generate the individual recommendation ranking list for a user, we first apply the **item-based top-n recommendation algorithm** (Deshpande & Karypis, 2004) to determine the items that are recommended to each user. Specifically, we compute the relevance between any two items r_i and r_j following the intuition: if more users like both items r_i and r_j, the relevance of r_i and r_j is higher. Then, for each user, we generate an individual recommendation ranking list based on the item relevance and **Borda count aggregation method** (Coppersmith et al., 2010). Briefly speaking, the user's individual list includes the items that have high relevance with her interested items marked in her profile. Afterwards, we still use the Borda count aggregation method to merge all the users' individual lists and get an aggregated recommendation ranking list for the community. Finally, we recommend the top K items from the aggregated list to all the community members.

Referring to the existing community recommendation work (Hu et al., 2014; Gorla et al., 2013), we exploit ***Average Precision*** to evaluate the recommendation performance for each community and compare ***Mean Average Precision*** over a set of selected communities to evaluate the effectiveness of CSD. Specifically, given a community $\mathcal{U}_c$ and an aggregated recommendation ranking list, assume that we recommend the top K items to a user $u \in \mathcal{U}_c$, we define the precision at rank position K for u as:

$$P@K(u) = \frac{rel_K(u)}{K} \tag{2}$$

where $rel_K(u)$ is the number of items that u likes among the top K recommendations. Then, for each community, we can calculate the Average Precision by:

$$AP@K(c) = \frac{1}{|\mathcal{U}_c|} \sum_{u \in \mathcal{U}_c} P@K(u) = \frac{1}{|\mathcal{U}_c|} \sum_{u \in \mathcal{U}_c} \frac{rel_K(u)}{K} \tag{3}$$

Finally, we define the Mean Average Precision for a given set $\mathcal{C}'$ of communities as:

$$MAP@K = \frac{1}{|\mathcal{C}'|} \sum_{c \in \mathcal{C}'} AP@K(c) = \frac{1}{|\mathcal{C}'|} \sum_{c \in \mathcal{C}'} \frac{1}{|\mathcal{U}_c|} \sum_{u \in \mathcal{U}_c} \frac{rel_K(u)}{K} \tag{4}$$

Note that we only use *precision*, but not *recall*, as the evaluation metric due to the following reason. In community selection for recommendation, intuitively we do not want to select the communities whose members only have few interested items (e.g., in movie recommendation, a community where the users generally dislike watching movies is absolutely not a good recommendation target). However, in the definition of recall, the denominator is the number of a user's interested items; thus, if a community has many users who have few interests, to some extent, it would be an advantage to get high recall, which contradicts the intuition. Therefore, recall is not an appropriate metric to evaluate the performance of community selection for recommendation.

Communities	$T_{C[1-100]}$	$T_{C[101-200]}$	$T_{C[201-400]}$	*All*	$B_{C[201-400]}$	$B_{C[101-200]}$	$B_{C[1-100]}$
Average CSD	0.052	0.012	0.009	0.005	0.0013	0.0011	0.0007
MAP@3	0.112	0.089	0.083	0.047	0.022	0.018	0.014
MAP@5	0.110	0.086	0.079	0.045	0.02	0.019	0.014
MAP@10	0.100	0.086	0.078	0.048	0.023	0.021	0.017

Table 1: MAP@K by CSD

6.2. Evaluation

We verify the effectiveness and efficiency of CSD in three experiments. (1) In the first experiment, we use CSD to select a set of communities and evaluate if the communities with larger CSD can generally gain better recommendation performance. (2) In Section 5.2, we have shown that Interest-based communities normally obtain larger CSD compared to the other types of communities; thus, in the second experiment, we investigate whether Interest-based communities can achieve higher recommendation precision. The second experiment evaluates the effectiveness of CSD in community selection from a different perspective compared to the first one. (3) In the last experiment, we study the computation time of CSD to reveal its efficiency for community selection.

6.2.1. Community Recommendation by CSD

Taking into account all the communities we introduced in Section 4.2, we respectively recommend top 3, 5 and 10 items to each community and compute the corresponding AP@K. Additionally, we generate a CSD ranking list by sorting all the communities based on their CSD in descending order. Then, we collect the successive communities in the CSD ranking list into various community sets and compare MAP@K of these community sets to examine whether recommending items to the communities with larger CSD can achieve higher precision.

We use $T_{C[n_1-n_2]}$ to represent the community set where the communities are in the top positions from n_1 to n_2 in the CSD ranking list; we use $B_{C[n_1-n_2]}$ to represent the n_1 to n_2 communities selected from the bottom of the CSD ranking list. Then, we expect that the community sets including the communities in the front of the CSD ranking list would achieve better recommendation performance (i.e., higher MAP@K) than the sets containing the communities in the back. Table 1 compares MAP@K among different sets of communities and verifies this expectation. Specifically, we observe that the community sets $T_{C[1-100]}$, $T_{C[101-200]}$, and $T_{C[201-400]}$ gain 7×, 3.5× and 3× larger MAP@K than the community sets $B_{C[1-100]}$, $B_{C[101-200]}$, and $B_{C[201-400]}$.

In addition, Figure 9 displays MAP@K of top N communities. The results show that the MAP@K declines with the increase of N. Note that, as N increases, more communities with smaller CSD are taken into account; thus, the average CSD of the top N communities decreases. In other words, Figure 9 indicates that when the average CSD of a community set decreases, the precision of recommendation for the set of communities also reduces.

Figure 10 plots CDF of AP@K of various community sets with different CSD. If we take the median (i.e., 0.5 in y-axis 'Percentage of Communities') AP@3 as an example, the results show that, selecting communities from the top 200 communities ($T_{C[1-200]}$) can achieve 1.5× larger AP@3 than random selection, and 5.7× larger AP@3 than selecting communities from the bottom 200 communities ($B_{C[1-200]}$).

In a nutshell, all the above experiment results demonstrate that selecting the communities with

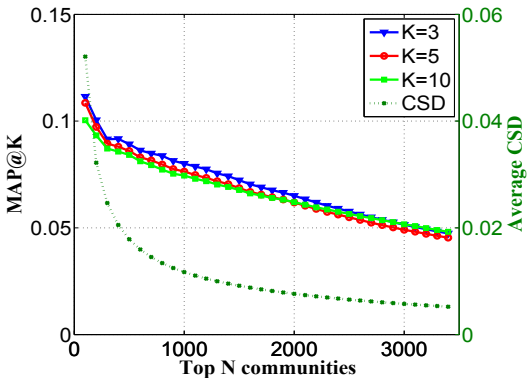
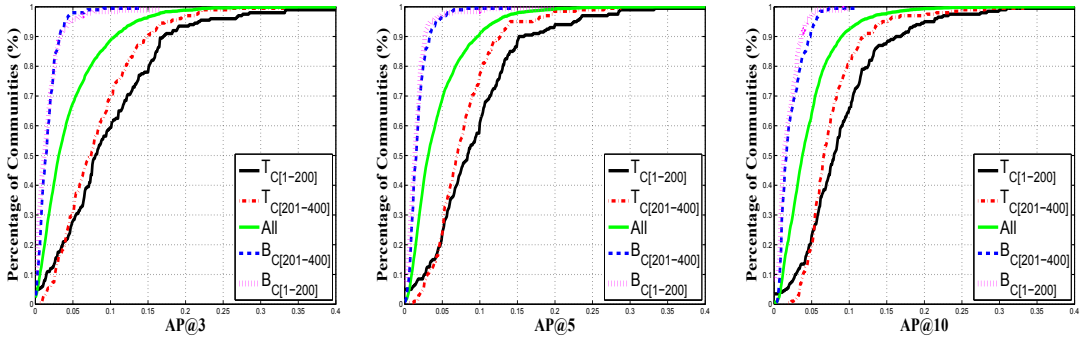


Figure 9: MAP@K of the top N communities in the CSD ranking list.



(a) K=3 (b) K=5 (c) K=10

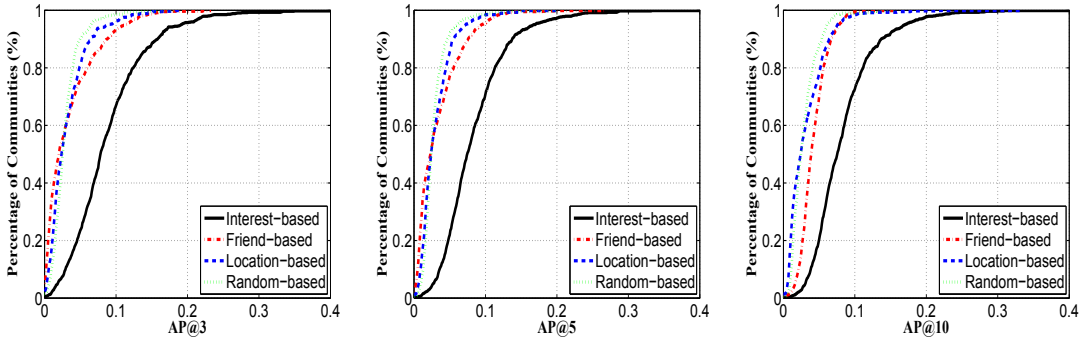
Figure 10: CDF of AP@K by CSD.

larger CSD can facilitate community recommendation with better performance. In other words, CSD can be used to select the appropriate communities to achieve higher precision in community recommendation.

6.2.2. Recommendation by Different Types of Communities

In this section, we evaluate the recommendation performance for various types of communities (Friend-based, Interest-based, Location-based and Random-based). As we have shown that Interest-based communities generally exhibit the largest CSD, we expect that the Interest-based communities can also achieve the highest MAP@K in community recommendation.

Table 2 and Figure 11 validate our expectation. Table 2 indicates that Interest-based community can produce about 2× MAP@K compared to the other three types of communities. Figure 11 plots CDF of AP@K for different types of communities. We observe the similar results that the median AP@K of Friend-based, Location-based and Random-based communities are all around 2× smaller than Interest-based communities. In addition, we notice that, although friendships are widely used to improve recommendation performance in much existing work (Purushotham et al., 2012; Yang et al., 2012), Friend-based communities do not perform as well as Interest-based communities. It may be because a user makes friends in various ways such as colleagues, families and classmates who do not necessarily share common interests with the user.



(a) K=3 (b) K=5 (c) K=10

Figure 11: CDF of AP@K by different types of communities.

Communities	Interest	Friend	Location	Random
Average CSD	0.0108	0.0036	0.0036	0.025
MAP@3	0.090	0.035	0.033	0.030
MAP@5	0.085	0.034	0.031	0.029
MAP@10	0.085	0.043	0.033	0.030

Table 2: MAP@K by different types of communities.

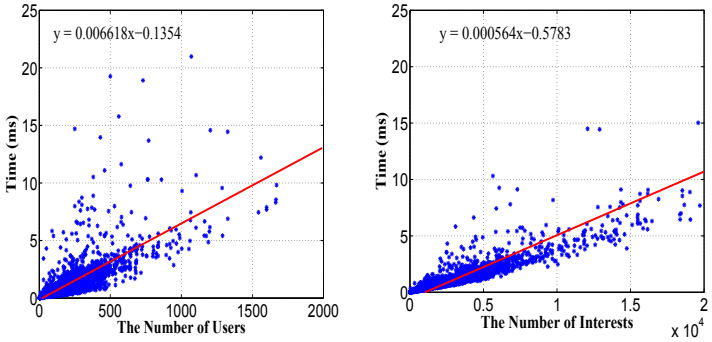
In summary, when only considering the community type, the results demonstrate that Interest-based communities are normally the best option for community recommendation. As the CSD of Interest-based communities is the highest among all four types (Section 5.2), these results also verify the effectiveness of CSD, i.e., the communities of higher CSD can achieve better recommendation performance.

6.2.3. Computation Time of CSD

In order to evaluate the efficiency of CSD for community selection, we record the time of computing CSD for all the four types of communities in our study (totally 3460 communities). By using an ordinary laptop (CPU: Intel Core i5-2540M 2.60 GHz; Memory: 6 GB; OS: Windows 7, 64-bit) and Python 2.7, the average computation time of CSD is 2.46 ms per community. According to this average speed, we can compute CSD for 1 million of communities within 41 minutes. In addition, Figure 12 shows the computation time versus the community size and the number of interests in a community, respectively. With regard to the linear regression models of computation time, it costs less than 10 ms on average to compute CSD for a community with 1500 users or with 20000 interests.

7. Conclusions

Community is fundamental and ubiquitous in various networks. For instance, biological functional communities build up and maintain metabolic networks; while social networks consist of groups of friends as well as various common location, interests and occupation based communities. So far, most network community studies have focused the effort on the techniques of detecting



(a) Community size (b) Number of interests

Figure 12: Computation time of CSD.

communities so as to facilitate certain applications; whereas, a huge number of self-organized communities have been present in Internet and real-life networks nowadays. In such circumstances, rather than detecting communities, evaluating the present communities and selecting the ones that can meet the specific requirements of applications are preferable. Specifically, this paper discusses the community selection for the application of community recommendation. To the best of our knowledge, this is the first work to discuss the research issue of quickly selecting the appropriate communities among a vast number of candidates in OSNs to improve community recommendation.

Taking users' interests in community recommendation as an instance, we have defined a metric of CSD to evaluate the interest similarity among users in a community. CSD indeed quantifies the inner connection density of community members by their common interests. In order to quickly estimate the interest similarity, we do not iteratively compute the similarities between all the user-pairs by using the conventional approaches (e.g., cosine similarity, Jaccard similarity or Pearson correlation coefficient) and then average the value. Instead, inspired by Lin's information-theoretic definition of similarity (Lin, 1998), we provide a formal definition of interest similarity among multiple users within a community. Specifically, the formulated metric CSD quickly divides the average popularity of interests by the total number of the community members.

In practice, apart from being used to improve community recommendation, CSD can serve various applications. Considering a sever deployment task in a content delivery network (CDN) which aims at selecting the best locations to deploy some new servers, we can first group the CDN users who live around a candidate location into a location-based community; Then, CSD can be employed to identify the best locations by selecting the best location-based communities in which the CDN users share the most similar interests.

According to different application requirements, the definition of CSD can be easily modified and extended. First, although CSD is originally defined for interest similarity estimation based on the community consisting of a set of users (U_c) and a set of interests (R_c), the interests of users can be replaced by other attributes. For instance, when it comes to a collaboration network, the set of users and the set of interests are replaced by a group of scientists and a set of collaborated publications/projects, respectively. Then CSD turns to assess the inner connection density of scientists inside a community by their collaborations. Second, given a certain attribute (e.g., collaboration, interest), CSD could be easily extended to a weighted CSD where a weight for each attribute instance needs to be considered in some specific applications. For instance, in a

collaboration network, larger weights can be put to more recent collaborations if we are concerned about the scientists' current collaboration status. To define the weighted CSD, only the popularity of each distinct attribute instance needs to be modified. In particular, for each distinct attribute instance, its popularity equals the product of the number of its fans and its weight.

Furthermore, CSD can be applied as a feature to help make intelligent decision (e.g., a feature in a machine learning algorithm). In a project examination and approval procedure, the quality of proposal, the strength of partners and the cooperation success degree of historical project may be all the determinants to select the qualified consortiums. Substituting interests and users in a community with historical projects and partners in a consortium respectively, CSD computes inner connection density of partners in terms of the degree of successful project cooperation. CSD then could work as a feature to select the qualified project applications.

Despite the above-mentioned efficiency in computation and effectiveness for various applications, in order to enhance CSD for more real-life applications, some issues still need study in the future.

First, the current definition of CSD has not considered the relatedness between different interests. For instance, the users who like *Harry Potter and the Philosopher's Stone* and the ones who are interested in *Harry Potter and the Chamber of Secrets* must have similarity in their interests; while CSD regards the two interests completely different and fails to consider their relatedness. Many approaches (e.g., cosine similarity, explicit semantic analysis, or latent semantic analysis) may be taken to compute CSD with consideration of the relatedness of different interests, while using these approaches may increase the computation complexity of CSD in turn. This opens a research issue for our future work — how to balance the computation complexity and the precision of interest similarity in real-life scenarios.

Second, CSD has not taken into account the popularity distribution of the interests inside a community. Let us look at two communities which both include 10 users and 2 interests. In the first community, 9 among the 10 users like the first interest and the rest user is a fan of the second interest; in the second one, half of the users like the first interest while the other half like the second interest. CSD assesses the interest similarity of both communities with the same value, even though the two communities have distinct interest popularity distributions that indicate different inner connection structures. At present, we believe it is a quite sophisticated task to ravel which one of the two communities has higher interest similarity among its users. Probably this will depend on the specific application that CSD is used to facilitate. We will start exploring this issue with some specific applications in the future work.

Acknowledgment

This work has been funded by China Scholarship Council and the project SITAC (ITEA2-11020). It has also been partially funded by the Ministerio de Economia y Competitividad of SPAIN through the project BigDatAAM (FIS2013-47532-C3-3-P) and the Program of National Natural Science Foundation of China (No. 41404025).

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 734–749.

Aimeur, E., Brassard, G., Fernandez, J. M., & Onana, F. (2006). Privacy-preserving demographic filtering. In *Proceedings of the 2006 ACM symposium on Applied computing* (pp. 872–878). ACM.

Backstrom, L., & Leskovec, J. (2011). Supervised random walks: predicting and recommending links in social networks. In *WSDM* (pp. 635–644). ACM.

Baltrunas, L., Makcinskas, T., & Ricci, F. (2010). Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of ACM RecSys* (pp. 119–126). Barcelona, Spain: ACM.

Basu Roy, S., Lakshmanan, L. V., & Liu, R. (2015). From group recommendations to group formation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 1603–1616). ACM.

Chen, J., Geyer, W., Dugan, C., Muller, M., & Guy, I. (2009). Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of ACM CHI* (pp. 201–210). Boston, MA, USA: ACM.

Chen, W.-Y., Zhang, D., & Chang, E. Y. (2008). Combinational collaborative filtering for personalized community recommendation. In *Proceedings of ACM SIGKDD* (pp. 115–123). Las Vegas, Nevada, USA: ACM.

Coppersmith, D., Fleischer, L. K., & Rurda, A. (2010). Ordering by weighted number of wins gives a good ranking for weighted tournaments. *ACM Trans. Algorithms*, *6*, 55:1–55:13.

Deng, S., Huang, L., & Xu, G. (2014). Social network-based service recommendation with trust enhancement. *Expert Systems with Applications*, *41*, 8075–8084.

Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems*, *22*, 143–177.

Dong, H., Hussain, F. K., & Chang, E. (2011). A service concept recommendation system for enhancing the dependability of semantic service matchmakers in the service ecosystem environment. *Journal of Network and Computer Applications*, *34*, 619–631.

Dougherty, J., Kohavi, R., Sahami, M. et al. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning: proceedings of the twelfth international conference* (pp. 194–202). volume 12.

Gjoka, M., Kurant, M., Butts, C., & Markopoulou, A. (2011). Practical recommendations on crawling online social networks. *JSAC*, *29*, 1872–1892.

Gorla, J., Lathia, N., Robertson, S., & Wang, J. (2013). Probabilistic group recommendation via information matching. In *Proceedings of WWW* (pp. 495–504). Rio de Janeiro, Brazil: ACM.

Han, X., Cuevas, A., Crespi, N., Cuevas, R., & Huang, X. (2014). On exploiting social relationship and personal background for content discovery in p2p networks. *Future Generation Computer Systems*, *40*, 17 – 29.

Han, X., Wang, L., Crespi, N., Park, S., & Cuevas, Á. (2015). Alike people, alike interests? inferring interest similarity in online social networks. *Decision Support Systems*, *69*, 92–106.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *28th Annual Meeting on Association for Computational Linguistics* (pp. 268–275). Association for Computational Linguistics.

Hu, L., Cao, J., Xu, G., Cao, L., Gu, Z., & Cao, W. (2014). Deep modeling of group preferences for group-based recommendation. In *Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 1861–1867). Pittsburgh, Pennsylvania, USA: AAAI.

Kusumoto, M., Maehara, T., & Kawarabayashi, K.-i. (2014). Scalable similarity search for simrank. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 325–336). ACM.

Lei, J.-B., Yin, J.-B., & Shen, H.-B. (2013). Gfo: a data driven approach for optimizing the gaussian function based similarity metric in computational biology. *Neurocomputing*, *99*, 307–315.

Lin, D. (1998). An information-theoretic definition of similarity. In *15th ICML* (pp. 296–304). Morgan Kaufmann Publishers Inc.

Makrehchi, M. (2011). Social link recommendation by learning hidden topics. In *Fifth RecSys* (pp. 189–196). ACM.

Markines, B., & Menczer, F. (2009). A scalable, collaborative similarity measure for social annotation systems. In *20th HyperText* HT '09 (pp. 347–348). ACM.

Masthoff, J. (2011). Group recommender systems: Combining individual models. In *Recommender Systems Handbook* (pp. 677–702). Springer.

Parameswaran, R., & Blough, D. M. (2007). Privacy preserving collaborative filtering using data obfuscation. In *IEEE International Conference on Granular Computing* (pp. 380–380). IEEE.

Purushotham, S., Liu, Y., & Kuo, C. (2012). Collaborative topic regression with social matrix factorization for recommendation systems. In *Proceedings of ICML* (pp. 759–766). Edinburgh, Scotland: Omnipress.

Quercia, D., Lathia, N., Calabrese, F., Di Lorenzo, G., & Crowcroft, J. (2010). Recommending social events from mobile phone location data. In *2010 ICDM* (pp. 971–976). IEEE Computer Society.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender Systems Handbook* (pp. 1–35). Springer US.

Ronen, I., Guy, I., Kravi, E., & Barnea, M. (2014). Recommending social media content to community owners. In *Proceedings of ACM SIGIR* (pp. 243–252). Queensland, Australia: ACM.

Rossi, R., Ahmed, N. K. et al. (2015). Role discovery in networks. *IEEE Transactions on Knowledge and Data*

Engineering, *27*, 1112–1131.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *10th WWW* (pp. 285–295). ACM.

Schwering, A. (2008). Approaches to semantic similarity measurement for geo-spatial data: A survey. *Transactions in GIS*, *12*, 5–29.

Spertus, E., Sahami, M., & Buyukkokten, O. (2005). Evaluating similarity measures: a large-scale study in the orkut social network. In *Proceedings of ACM SIGKDD* (pp. 678–684). Chicago, Illinois, USA: ACM.

Tao, W., Yu, M., & Li, G. (2014). Efficient top-k simrank-based similarity join. *Proceedings of the VLDB Endowment*, *8*, 317–328.

Tsebelis, G. (1995). Decision making in political systems: Veto players in presidentialism, parliamentarism, multicameralism and multipartyism. *British journal of political science*, *25*, 289–325.

Yang, X., Steck, H., & Liu, Y. (2012). Circle-based recommendation in online social networks. In *Proceedings of ACM SIGKDD* (pp. 1267–1275). Beijing, China: ACM.

Zhang, J., Tang, J., Ma, C., Tong, H., Jing, Y., & Li, J. (2015). Panther: Fast top-k similarity search on large networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1445–1454). ACM.