We Know Who You Are: Discovering Similar Groups Across Multiple Social Networks

Xiaoming Liu, Chao Shen, *Member, IEEE*, Xiaohong Guan, *Fellow, IEEE*, and Yadong Zhou, *Member, IEEE*

***Abstract*—People use various online social networks for different purposes. The user information on each social network is usually partial. Thus, matching the users across these multiple online social networks is of great significance for providing new services as well as new insights on user behaviors. Recent research shows that group structure widely exists on social networks, in which members work together with certain purpose and are more influential than individuals on online social networks. Previous works provide outstanding solutions for mapping individuals, but few ones pay attention to the study of groups across multiple social networks. To address the research gap, we first aim to propose an effective method to detect similar group across multiple social networks. The method mainly has three steps, including detecting group structure based on random walks, extracting similarity features, and inferring group similarity using probabilistic graphical model. We evaluate our algorithm on five different types of online social networks. Experimental results show that our method achieves 0.693, 0.779, 0.729 in precision, recall, and *F*1-measure, which significantly surpass the state-of-the-art by 31.4%–44.3%, 17.3%–26.3%, and 25.9%–31.6%, respectively. The outstanding performance of our method demonstrates that our proposal can reach the requirement of detecting similar groups across social networks. In particular, the result of this paper paves the way for the recommendation system, link prediction and information diffusion across sites.**

***Index Terms*—Algorithm, experiments, group similarity, multiple online social networks, probabilistic graphical model, random walks.**

I. Introduction

ONLINE social networks play an important role in people's daily lives with different types of forms, such as tweets, business, music sharing, game services, and blogs, etc [1]. A report[1] explicates that the number of global users would reach about 2.95 billion by 2020, around one-third of the world's population. People sigh up multiple social network accounts for these different purposes. For example, users mainly communicate with friends by creating accounts on Facebook, share interest groups through using Google+, and estimate professional networks on Linkedin, etc [2]. A simple survey by the researchers indicates that the average number of online social networks that one user joins is eight [3].

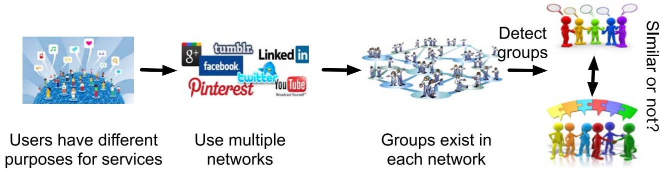


Fig. 1. Brief introduction of discovering similar groups across multiple online social networks.

Recent research reveals that user collaboration with group structure exists widely in different kinds of online social networks [4], such as spam information diffusion [5] and fake customer reviews [6]. Thus, discovering similar groups across multiple online social networks is of great significance to the study on group behaviors. For example, many companies, such as PeekYou[2] and Spokeo,[3] offer the services of people search by aggregating the public visible user information from multiple online social networks; what is more, recommender service provider, like Yelp and Amazon, can filter out spammer reviews by linking their accounts and analyzing their behaviors on other online social networks [7].

In this paper, the compared groups are from different online social networks, and similar groups would share similar users and have similar group structures. We briefly introduce our problem in Fig. 1 about discovering similar groups across multiple online social networks. As shown in Fig. 1, users have different purposes for Web services in their daily life. In order to meet their purposes, they take part in multiple online social networks, such as Facebook, Google+, and Linkedin, etc. While, group structure exists in each network, and we need to detect this groups in each network. By calculating the similarity of the user relationships and social contents in these two groups from two different networks in the virtual-network world, we attempt to judge whether they are just the same group of people or not in the real-life world.

Proposing this problem from the perspective of groups is inspired by several motivations. First, group structure [8] is proved to widely exist in various social networks [4].

Manuscript received March 1, 2018; accepted April 3, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61403301, Grant 61773310, Grant 61572397, Grant U1766215, Grant 61502383, and Grant U1736205, and in part by the Open Project Program of the National Laboratory of Pattern Recognition. This paper was recommended by Associate Editor Z.-G. Wu. *(Corresponding author: Chao Shen.)*

The authors are with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: xmliu@sei.xjtu.edu.cn; cshen@sei.xjtu.edu.cn; xhguan@sei.xjtu.edu.cn; ydzhou@sei.xjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2018.2826555

[1] https://www.statista.com/topics/1164/social-networks/

[2] https://www.peekyou.com/

[3] https://www.spokeo.com/

Recently, researchers also show that spam information diffusion on online social networks [5] and fake customer in product reviews [6] conduct their benefit-oriented actions in groups. What is more, groups usually have much larger influence than individual users in information diffusion on online social networks. Detecting these groups from multiple social networks has far-reaching implications for the research of online social networks. Second, based on the wealthy information of these detected similar groups across multiple social networks, we can do social media mining [9], [10] more accurately. Though each site may only have limited information about one group, the complementary information for this group can be provided by the other networks. But in order to combine the useful information to build better group features, we have to reliably link the similar groups across online social networks. Third, difficulties of fragmentary data for the single user can be solved by the information complementarity in groups. Limited by the permission restriction set by online social networks or user, it is hard to obtain the user personal information completely, which may be a big challenge for the mapping of the individual user across multiple online social networks. But if we change the mapping object from the individual user to the group, the problem of incomplete data could benefit from the rich information of the group information. The various kinds of features in one group could help eliminate some noise in individual information across different social networks, i.e., some users may have different profiles and behavior patterns on different online social networks.

Although this significant problem seems to be promising, it is faced with many challenges. First, with the scale of social media data becoming larger and larger, the problem of information missing for certain users is hard to be avoided in the study of user similarity. This obstacle is led by several reasons, such as access denied or not be submitted. Therefore, how to discover these similar groups in these very large networks with scattered information is extremely challenging. Second, high computational complexity is always a critical challenge in the analysis of big data, especially for big graphs, which are very difficult to be parallelized. Discovering the group structure is a vertex clustering problem in graphs, which has been proved to be an NP-complete problem [11]. Besides, to calculate the similarity of two groups across different online social networks, it also costs lots of computational resources. Here, we present a simple analysis. Let G_1 and G_2 denote two networks, and n_1 and n_2 represent the number of users in these two networks. In theory, the time complexity of computing the similarity value for all these group pairs is not lower than $O(n_1 n_2)$. Third, labeling ground truth is one of the most challenging obstacles in the user mapping research across online social networks.

To address the research gap and technique challenges, we model the similar group detection across multiple online social network problem systematically and mathematically. Taking advantage of this model, we propose an efficient method to discover these similar groups across online social networks. This method mainly consists of three steps: first, we detect groups in each network using random walks, respectively; second, we extract six similarity features of each two groups from the views of group profiles and group structure; third, we take these similarity features as input to train the probabilistic graphical model, which is used for inferring the two groups are similar or not. We also detail the rule of labeling ground truth. We evaluate our method on the five different online social networks. Compared with the state-of-the-art, our method shows the advantage in dealing with these challenges of this problem. We also discuss the scalability of our algorithm to solve the challenge of high time complexity and the possible application scenarios. For revealing the significance of this paper intuitively, we take link prediction as a case study and show the improvement benefited from our result.

The research in this paper takes a step toward group identifying with an effective and low-cost way by mapping similar groups across multiple online social networks, which enhances the application of social networks, such as user migration, friend recommendation, information diffusion, and network dynamic analysis [12]. We summarize our major contributions as follows.

1) We propose a novel problem of discovering similar groups across multiple online social networks, which would enhance the performance of various kinds of social network applications. The study of mapping problem across social networks from the perspective of the group also verify their potential advantages.
2) We model this problem systemically and mathematically, including employing random walks to detect groups in social networks, extracting the similarity features of two groups, and inferring whether two groups are similar or not by the probabilistic graphical approach. Based on this model, we propose an efficient algorithm to discover similar groups across multiple social networks. We also propose solutions to deal with the technical challenges of this mapping problem, including incomplete data, high time complexity, and the determination of ground truth.
3) We evaluate our algorithm on five different large-scale online social networks from the real-life world. The metrics of our detected similar group results achieve 0.693, 0.779, and 0.729 in precision, recall, and $F1$-measure, respectively. We also compare our algorithm with the state-of-the-art, which shows that ours surpass the competitors 44.3% at most. We also present the scalability of our method varying with the number of computer nodes.
4) We take advantage of these detection result to improve the performance of link prediction problem, which demonstrates the significant performance gains with respect to the $F1$-measure.

The organization of this paper is as follows. Section II shows the definition of this problem. Section III presents the similar group detection model based on the random walks and probabilistic graphical model to infer the similarity of two groups from different networks. Section IV gives the experimental evaluation on five different large-scale online social networks. Section V discusses the scalability. Section VI summarizes the research related to this paper. The scope is concluded in Section VII with conclusions and future work.

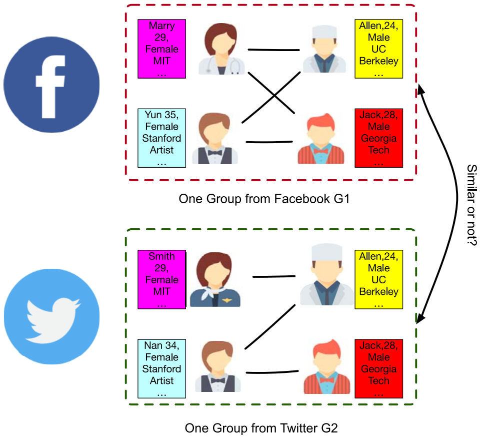

Marry 29, Female MIT ...
Allen,24, Male UC Berkeley ...
Yun 35, Female Stanford Artist ...
Jack,28, Male Georgia Tech ...
One Group from Facebook G1
Smith 29, Female MIT ...
Allen,24, Male UC Berkeley ...
Nan 34, Female Stanford Artist ...
Jack,28, Male Georgia Tech ...
One Group from Twitter G2
Similar or not?

Fig. 2. Illustration of discovering similar groups across social networks.

II. Problem Definition

Let Gs denote the multiple online social networks, which are constructed by different social networks $G_1, G_2, \ldots, G_k$. Each network contains user contents and relationships, as shown in Fig. 2. Suppose that g_1 and g_2 are two groups from two different online social networks G_i and G_j, respectively, where $G_i, G_j \in Gs$. What we want to do is that by comparing the users in group and group structure, the algorithm would judge whether g_1 and g_2 being made up of similar users or event the same users with certain relationships. We divide this problem into two categories: one case is that we have no user mapping relationship result in these two groups, except for the user profiles and relationship topology; another case is that we know both partial mapping relationship of users in two groups and the user profiles with relationship topology. In this paper, we do our best to solve the first one problem which is more challenging than the second one, as shown in Fig. 2. Through this paper, graph and network are considered to be the same.

With the formulation introduced above, we present the definition of our problem as follows.

Input: We take k different online social networks $G_1, G_2, \ldots, G_k$, which is denoted by GS, as input. For $G_i \in GS$, $G_i = \{V_i, E_i, P_i\}$ is one network, where V_i represent the users in this network, E_i denotes the relationships among users, and P_i is the profile content for these users.

Output: For any two groups from any two different networks, i.e., $g_1 \in G_i$ and $g_2 \in G_j$, the proposed algorithm is required to discover them and infer whether g_1 and g_2 are similar groups or not in the real-life world.

III. Methodology

The framework of our method is shown in Fig. 3. After inputting the multiple online social networks, we use random walks to detect the groups in each network first. Then, we extract the similarity features for each pair of groups from different networks in different perspectives, including the group profile and relationship structure. Probabilistic graphical model, which is constructed based on training data, are used for inferring whether two groups are similar or not finally.

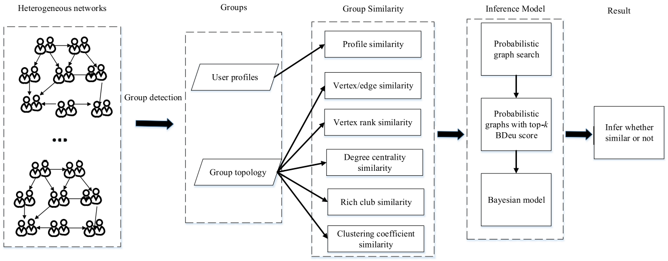


Fig. 3. Framework of our method.

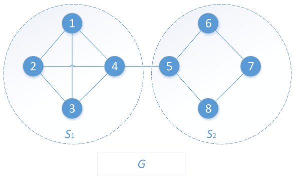


Fig. 4. One graph with two subgraphs for analysis.

A. Detecting Groups in Networks

One group is made up of two or more closely connected users who are with certain cooperation relationship or friendship, and share similar characteristics or interests. Random walks is always used for exploring the graph structure in the field of sampling and estimating networks [13]. In this paper, it is also employed to discover group structure in multiple networks. Because random walks is easily trapped in the closely connected subgraphs, which is the group structure that we want to discover. Making use of this nature of random walks, we can detect the groups effectively. What is more, the theory of Markov chain and graph spectral also support the mathematical foundation for the solidity of the proposal. We introduce our basic idea to discover these subgraphs containing closely connected vertices in networks below.

1) Inspiration: Fig. 4 is an example of graph for analysis, which have two subgraphs S_1 and S_2. S_1 has more paths than S_2. In order to simplify the analysis, we just consider one special type path (three edges and four different vertices). S_1 has 24 different such special paths, while S_2 just has 8 different such ones. Inspired by this distinction, we propose an indicator to reveal the closeness of the vertices.

Our basic idea is that we consider both the edges traversed by the random walks and the other edges being between any vertex pair in the path. Avoid misleading the readers, we detail the difference between "the computed edges" and "the traversed edges." For instance, if random walks traverse a path $\{v_1, v_2, v_3, v_4\}$, the edges e_{12}, e_{23}, e_{34} are traversed edges, and both the traversed edges and the other edges e_{14}, e_{13}, e_{24} which exist between vertex pairs in this path are computed edges.

Suppose that the random walks process is naive. Based on the theory of random walks, the paths in S_2 is less likely to be traversed by the walker than the paths in S_1, because the vertices in S_1 have the larger degrees. The probability of one such path in S_1 traversed by the random walks is denoted by p_1, and p_2 for S_2, where $p_1 > p_2$. In this situation, the probability of an edge in S_1 being computed is $24p_1$, but it is $8p_2$ in S_2. Obviously, the probability of computing an edge in

S_1 with more closely connected vertices is much larger than that in S_2.

In order to present more general conclusion, we take two cases study for the number of such special paths, which consists of $n_0 - 1$ edges and all the n_0 different vertices in the subgraph. Let W_1 denote the number of such special paths in complete subgraph in which vertices are connected closely, and W_2 for the amount of such special paths in cycle subgraph in which vertices are almost the most sparsely connected. Through analyzing, we can get that $W_1 = n_0!$, $W_2 = 2n_0$, and

$$\frac{W_1}{W_2} = \frac{(n_0 - 1)!}{2}. \tag{1}$$

From (1), we can see that such special paths in complete subgraph are much more than those in cycle subgraph. Let the probability of one of such paths traversed by random walks in one complete subgraph being denoted by p'_1, and the probability in one cycle subgraph is p'_2, where $p'_1 > p'_2$. In this situation, we can obtain the probability of computing an edge in complete subgraph, which is $W_1 p'_1$, while in cycle subgraph the probability is $W_2 p'_2$. When the random walks arrives its stationary probability distribution, the number of each computed edge is about $tW_1 p'_1$ in complete subgraph, while, in cycle subgraph there is just about $tW_2 p'_2$, where t is the number of steps. The gap of the edges being computed between these subgraphs with close relationship and subgraphs with sparse connection is bigger than $[(n_0 - 1)!/2 - 1]tp'_2$. We take advantage of this idea to deal with the problem of group detection.

2) Group Detection Based on Random Walks: A network is denoted by $G_i = (V_i, E_i)$, where V_i represent the vertex set (users) and E_i is the edge vertex (relationships). Inspired by the feature of random walks, we propose an algorithm to detect groups in each network. The initial vertex v_0 of random walks is selected from the vertex set V_i randomly with the same probability distribution. The walk process terminals when its number of hops reaches its budget $10n\sqrt{n}$, where n is the number of vertices. The vertex sequence obtained by random walks is

$$\mathrm{VS} = \left\{v_a, v_b, \ldots, v_q \mid v_a \in V_i \wedge v_b \in V_i \wedge \ldots \wedge v_q \in V_i\right\}. \tag{2}$$

After finishing random walks, the long path is spitted into shorter paths with length k by k-splitting strategy [14]. Therefore, we get a new sequence VSS with short paths

$$\mathrm{VSS} = \left\{\left\{v_a, \ldots, v_f\right\}, \ldots, \left\{v_c, \ldots, v_q\right\}\right\} \tag{3}$$

where $\{v_a, \ldots, v_f\}, \ldots, \{v_c, \ldots, v_q\}$ are the shorter paths with length k.

Then we weight the edges by computing the existing edges among the vertices in these short paths. $Ne(v_i, v_j)$ represents the times of e_{ij} being computed. Its set is

$$\mathrm{NeS} = \left\{Ne(v_d, v_a), Ne\left(v_f, v_u\right), \ldots, Ne\left(v_i, v_j\right)\right\}. \tag{4}$$

When we finish weighting edges, we sort them in descending order by their weight. The sorted edges is stored in list which is denoted by

$$\mathrm{SE} = \left\{e_{df}, \ldots, e_{sh} \mid e_{df} \in E \wedge, \ldots, \wedge e_{df} \in E\right\}. \tag{5}$$

Based on the sorted edge sequence, we detect group structure in the network G_i as following rules. First, we initialize the first group g_0 by the two vertices in the largest weighted edge. For the coming edge in SE, if g_0 has one of these two vertices, then another vertex is added to g_0; if both of the two vertices are not in g_0, then they would make up of a new group g_1. According to these rules, the edges in SE are assigned to groups one by one. Finally, we obtain the group structures in each network of these multiple online social networks.

B. Similarity Features of Groups Across Multiple Online Social Networks

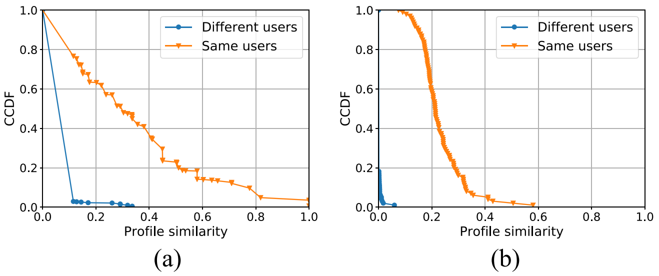
Suppose that g_1 and g_2 are two groups from two different networks G_i and G_j, respectively. We extract the similarity features of these two groups from different kinds of perspectives, including the profile similarity and topology similarity of groups.

1) Profile Similarity: The information of the user profile presents the basic description of one user. For example, username gives the information what we call him, from location information we could obtain where the user lives in, link provides the more detail information for the user by his homepage, hometown tells us where he is born, age shows that how old he is, and status indicates that whether he is in relationship or married, etc. Previous work proved that the similar users across multiple different online social networks should share similar user profile [3]. Instead of taking advantage of all the profile information, we adopt three factors, i.e., the username, location, and homepage link. Because these description for one user usually have obviously personal characters, which can distinguish him from the other users. What is more, these items are usually similar across multiple social networks, which can help us discover them across different online social networks. We split these strings for one user into one word bag, and employ TF-IDF algorithm (term frequency-inverse document frequency) to process the word bad and obtain its weighted vector. We would like to use weighted vector $\mathbf{w}_{g_1}$ and $\mathbf{w}_{g_2}$ to denote the TF-IDF results for g_1 and g_2. Finally, we compute the profile similarity value by making use of the cosine distance based on these two vectors as follows:

$$\mathrm{sim_p}(g_1, g_2) = \frac{\mathbf{w}_{g_1} \cdot \mathbf{w}_{g_2}}{||\mathbf{w}_{g_1}||||\mathbf{w}_{g_2}||}. \tag{6}$$

For verifying the solidity of our proposal, we take one network pair, i.e., Flickr and Last.fm, as a case study and present the measurements on the user profile similarity for these two networks, as shown in Fig. 5. Fig. 5(a) is the complementary cumulative distribution of the profile similarity based on the selected profile (username, location, and homepage link). Fig. 5(b) is the measurement based on the all profile information. These two figures both show that there is a big difference between the same user profile similarity value and the different user similarity value. But by filtering the useless information, we can reduce lots of computational resource cost.

2) Topology Similarity: The user profile is not available for everyone. To deal with the challenge of incomplete data and



(a) (b)

Fig. 5. Complementary cumulative distribution function of similarity values for the same users and different users from the two different networks Flickr and Last.fm. (a) Selected profile information. (b) All profile information.

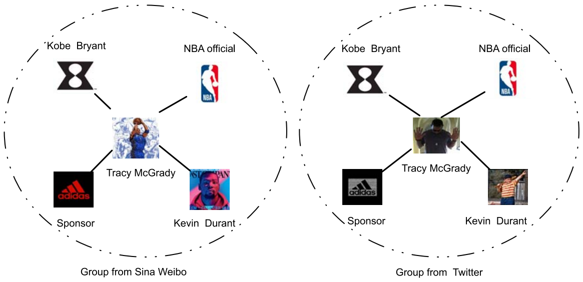


Fig. 6. Example of topology structure of Tracy McGrady in Sina Weibo and Twitter.

the potential problem of only taking account of the user profile, the group topology structure is also adopted to compare the groups in multiple online social networks. As we describe our problem earlier, two similar groups from different social network platforms should share many similar users and have the same type of relationship structure.

In order to give a more direct explanation for this phenomenon, we illustrate it by an example of the main relationship for the groups of one famous NBA star Tracy McGrady in Sina Weibo[4] and Twitter in Fig. 6. From Fig. 6 we can see that he has followed his sponsor, the NBA official account, and the other NBA famous stars in both of the social networks. It reveals that the relationships between McGrady and his friends are quite similar in different online social networks.

To conduct a deeper analysis of this problem, we classify the users on online social networks into two categories based on their degree. They are leader users who are with the large degree and common users who have the small degree. For example, McGrady has more than 4 million fans on Sina Weibo, but just follows tens of famous stars or related official accounts. Thus, Tracy McGrady and his fans make up of one group with the star-topology, and McGrady is one of the leader users. This star-topology is very common in the groups of the other famous stars, such as Allen Iverson and Kobe Bryant. In this situation, even if we have no information about the user in place of McGrady in the Twitter group, considering his social position, we could infer his identity maybe McGrady through comparing the information of other users in Twitter groups and Sina Weibo group and the topology of these two groups.

Let $V_{g_1}, V_{g_2}, E_{g_1}$, and E_{g_2} denote the vertex sets and edge sets of groups g_1 and g_2. The group similarity is calculated based on graph similarity theory.

The vertex/edge overlap (veo) distance is used for measuring the topology similarity by finding the similar users and relationships in two groups

$$\mathrm{sim}_{\mathrm{veo}}(g_1, g_2) = 2\frac{|V_{g_1} \cap V_{g_2}| + |E_{g_1} \cap E_{g_2}|}{|V_{g_1}| + |V_{g_2}| + |E_{g_1}| + |E_{g_2}|}. \tag{7}$$

Another topology similarity of these two groups is based on the vertex score obtaining by PageRank algorithm [15]. PageRank algorithm achieves outstanding performance for valuing the importance of one Web page. We take advantage of it to evaluate the social character of one user in one group. If the users from two groups have similar social character distribution, it would increase the credit that these two groups are similar. The score of $v_k \in V_{g_1}$ is calculated by

$$\mathrm{PR}(v_k) = \frac{1-d}{|V_{g_1}|} + d\sum\nolimits_{v_m \in M(v_k)} \frac{\mathrm{PR}(v_m)}{L(v_m)} \tag{8}$$

where d denotes the damping factor, $M(v_k)$ represents the vertices pointing to v_k, and $L(v_m)$ is the number of vertices which are pointed by v_m. Thus, the vertex rank (vr) distance is defined as follows:

$$\mathrm{sim}_{\mathrm{vr}}(g_1, g_2) = 1 - \frac{2\sum\nolimits_{v \in V_{g_1} \cup V_{g_2}} w_v\left(\pi_v^1 - \pi_v^2\right)}{D} \tag{9}$$

where π_v^1 and π_v^2 are the ranks of v in the sorted list based on PageRank scores for g_1 and g_2, $w_v \sim d_v$ is the quality of v, and D is a normalization factor that limits the maximum value of the fraction to 1.

We also consider three graph features to measure the similarity of two groups, including degree centrality, rich club coefficient, and clustering coefficient.

Degree centrality (dc) for a vertex is the fraction of vertices that the vertex is connected. The degree centrality values are normalized by dividing the possible maximum degree $n-1$, where n is the number of vertices in graph. It reveals the degree distribution of one group. The degree centrality list is denoted by

$$DC = [dc_1, dc_2, \ldots, dc_n]. \tag{10}$$

Note that values in DC is sorted in descending order. Suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two vectors of degree centrality of group g_1 and group g_2, their similarity is calculated by their correlation coefficient

$$\mathrm{sim}_{\mathrm{dc}}(g_1, g_2) = \frac{\mathsf{E}[\boldsymbol{XY}] - \mathsf{E}[\boldsymbol{X}]\mathsf{E}[\boldsymbol{Y}]}{\sqrt{\mathsf{E}[\boldsymbol{X}^2] - \mathsf{E}[\boldsymbol{X}]^2}\sqrt{\mathsf{E}[\boldsymbol{Y}^2] - \mathsf{E}[\boldsymbol{Y}]^2}} \tag{11}$$

where E is the expectation. If two groups do not have the same number of vertices, we would add zeros at the bottom of the short degree centrality vector.

The function of one group is usually determined by the users who have more friends. Rich club (rc) coefficient is the ratio of the number of edges among the vertices whose degrees are

[4] http://weibo.com

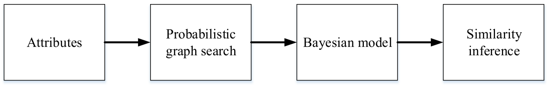


Fig. 7. Work-flow of inference model for group similarity.

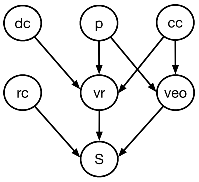


Fig. 8. Illustration of a probabilistic graph.

larger than k. It indicates the relationship between leader users in one group. It is defined as

$$\phi(k) = \frac{2E_k}{N_k(N_k - 1)} \tag{12}$$

where N_k is the number of vertices with degree larger than k, and E_k is the number of edges among these vertices. Because the rich clubs of two groups may have large differences in degree values, the value of k for each group takes $0.1 * d_{\max}$, where $d_{\max}$ is their largest degree value in its group. In this case, if two groups have similar leader relationship topology, two groups should have similar rich club coefficient, even with large differences in degree values. Suppose that ϕ_1 and ϕ_2 are the coefficients of two groups, the rich club similarity is calculated by

$$\mathrm{sim}_{\mathrm{rc}}(g_1, g_2) = 1 - |\phi_1 - \phi_2|. \tag{13}$$

Clustering coefficient (cc), which shows the closeness of user connections in one group, is also applied to measure the similarity of groups

$$\mathrm{CC} = \frac{1}{n_g}\sum\nolimits_{v \in g} \frac{2T_v}{d_v(dv - 1)} \tag{14}$$

where g is one group, T_v is the number of triangles through vertex v, d_v is the degree of v, and n_g is the number of vertices in group g. Their similarity can also be calculated by (13).

C. Inference Using Probabilistic Graphical Model

The workflow of our inference model is shown in Fig. 7, which is used for deciding whether two groups are similar or not. We take the similarities introduced in Section III-B as the input attributes. By maximizing BDeu score, we discover the causal relationships among the six similarities and group similarity result, which is denoted by probabilistic graphs [16]. These graphs are used for building the Bayesian network model [17] to predict the probability of the two groups being similar.

Probabilistic graphical model uses a graph to express the conditional dependence structure between random variables. In one probabilistic graph, the vertices represent the variables, and edges denote the causal relationship between variables. Its illustration is shown in Fig. 8. The edge in Fig. 8 from vertex v_{p} to vertex v_{veo}, i.e., $(v_{\mathrm{p}}, v_{\mathrm{veo}})$, means that variable $\mathrm{sim}_{\mathrm{p}}$ (group profile similarity) has a causal effect on variable $\mathrm{sim}_{\mathrm{veo}}$ (the similarity of overlapping vertex and edge).

1) Bayesian Network Model: In our problem, the useful information is very sparse and lots of the data is incomplete. Thus, the scale of training data set is limited and many of them are uncertain. Bayesian network model using multiple probabilistic graphs can obtain conclusions which are not susceptible to incorrect categorical decisions about independence facts that can occur with data sets of finite size. What is more, the Bayesian network approach can discover good causal relationships and make good inferences for modeling uncertainty based on the multiple probabilistic graphs. Therefore, Bayesian network approach meets our needs and can be used in our inference problem.

Here we present the prediction problem based on Bayesian networks in mathematic format. $X = \{\mathrm{sim}_{\mathrm{p}}, \mathrm{sim}_{\mathrm{veo}}, \mathrm{sim}_{\mathrm{vr}}, \mathrm{sim}_{\mathrm{dc}}, \mathrm{sim}_{\mathrm{rc}}, \mathrm{sim}_{\mathrm{cc}}, \mathrm{sim}_{\mathrm{g}}\}$ is the inference problem variable set, where $\mathrm{sim}_{\mathrm{g}} = \{\mathrm{yes}, \mathrm{no}\}$ is the inference variable of whether two groups being similar or not, and the description of other variables is in Section III-B. $D = \{d_1, d_2, \ldots, d_n\}$ is the data set which is randomly sampled for the unknown probability distribution for X. For each case $d_i \in D$, there is an observation of all variables in X. T is the probabilistic graph set variable whose state t is the possible true probabilistic graph with probability distribution $p(t)$.

Let h denotes the hypothesis of next observation for X, i.e., one possible situation of two groups are similar. Given the sampled data D, the probability that the hypothesis is true is

$$p(h|D) = \sum\nolimits_{t \in T} p(t|D)p(h|D, t) \tag{15}$$

where $p(t|D)$ is the probability that the probabilistic graph is true given sampled data D, and $p(h|D, t)$ is the probability that the hypothesis is true given the sampled data D and probabilistic graph pg. We can calculate $p(t|D)$ by

$$p(t|D) = \frac{p(t)p(D|t)}{\sum_{\hat{t}} p(\hat{t})p(D|\hat{t})} \tag{16}$$

where $p(\hat{t})$ is usually assumed to be a uniform prior distribution, i.e., $p(\hat{t}) = (1/|T|)$.

Based on the Causal Markov condition [18], the BDeu score is

$$p(D|t) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{17}$$

where n is the number of vertices in probabilistic graph, α is the equivalent sample size, $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. F_i denotes the parent vertices of vertex V_i in probabilistic graph, and its configuration is f_i. Variable r_i is the number of possible values of V_i whose configuration is v_i, so F_i has $q_i = \prod_{V_j \in F_i} r_j$ possible values. N_{ijk} is the number of cases in D where $V_i = v_i^k$ and $F_i = f_i^j$. For any complete probabilistic graph t_c, we have [19]

$$\alpha_{ijk} = \alpha p\left(v_i^k, f_i^j \middle| t_c\right) \tag{18}$$

where $p(v_i^k, pa_i^j|pg_c)$ is computed from the joint probability distribution $p(X|pg_c)$. For the incomplete probabilistic graph, we can make use the assumption of *parameter modularity* [20] which says that if V_i has the same parents in probabilistic graph pg_1 and pg_2, then $\alpha^1_{ijk} = \alpha^2_{ijk}$. For an arbitrary probabilistic graph, we can compute α_{ijk} by constructing a complete one. For example, if vertex V_i has parents Pa_i in one given model, we can compute α_{ijk} based on one complete structure where V_i has these parents by (18). Suppose h is the next data case d_{n+1}, then

$$p(h|D, m) = \prod_{i=1}^{n} \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}. \tag{19}$$

When two groups g_1, g_2 are detected from the multiple online social networks, their similarity observation values are $\{\mathrm{sim}^{1,2}_{\mathrm{p}}, \mathrm{sim}^{1,2}_{\mathrm{veo}}, \mathrm{sim}^{1,2}_{\mathrm{vr}}, \mathrm{sim}^{1,2}_{\mathrm{dc}}, \mathrm{sim}^{1,2}_{\mathrm{rc}}, \mathrm{sim}^{1,2}_{\mathrm{cc}}\}$. Let $h^{1,2}_{\mathrm{yes}}$ denote the hypothesis of the situation that these two groups are similar, i.e.,

$$h^{1,2}_{\mathrm{yes}} = \left\{\mathrm{sim}^{1,2}_{\mathrm{p}}, \mathrm{sim}^{1,2}_{\mathrm{veo}}, \mathrm{sim}^{1,2}_{\mathrm{vr}}, \mathrm{sim}^{1,2}_{\mathrm{dc}}, \mathrm{sim}^{1,2}_{\mathrm{rc}}, \mathrm{sim}^{1,2}_{\mathrm{cc}}, \mathrm{yes}\right\}$$

and let $h^{1,2}_{\mathrm{no}}$ represent the hypothesis of the situation that these two groups are not similar, i.e.,

$$h^{1,2}_{\mathrm{no}} = \left\{\mathrm{sim}^{1,2}_{\mathrm{p}}, \mathrm{sim}^{1,2}_{\mathrm{veo}}, \mathrm{sim}^{1,2}_{\mathrm{vr}}, \mathrm{sim}^{1,2}_{\mathrm{dc}}, \mathrm{sim}^{1,2}_{\mathrm{rc}}, \mathrm{sim}^{1,2}_{\mathrm{cc}}, \mathrm{no}\right\}.$$

We use equations (15)–(19) to calculate $p(h_{\mathrm{yes}}|D)$ and $p(h_{\mathrm{no}}|D)$. Then given the sampled data D, the probability of the two groups being similarity on the condition $con = \{\mathrm{sim}^{1,2}_{\mathrm{p}}, \mathrm{sim}^{1,2}_{\mathrm{veo}}, \mathrm{sim}^{1,2}_{\mathrm{vr}}, \mathrm{sim}^{1,2}_{\mathrm{dc}}, \mathrm{sim}^{1,2}_{\mathrm{rc}}, \mathrm{sim}^{1,2}_{\mathrm{cc}}\}$ is

$$p\big(\mathrm{sim_g} = \mathrm{yes}|con, D\big) = \frac{p\big(h_{\mathrm{yes}}|D\big)}{p\big(h_{\mathrm{yes}}|D\big) + p(h_{\mathrm{no}}|D)}. \tag{20}$$

This probability means the possibility of two groups are similar in this condition.

IV. Evaluations

In this section, we introduce the setting of our experiments, and present the comparison between our method and the state-of-the-art. We analyze the feature contribution and the efficiency of our method. In order to show the improvement of this paper to other applications, we take link prediction as one case study.

A. Experimental Setting

The data comes from COSNET [3] which is a research paper focuses on connecting similar users in heterogeneous networks. This dataset provides both user relationships and profiles in five different online social networks, whose description is shown in Table I. User profiles, which provides the basic but useful information of users, are very sparse.

Our ground truth is based on the user pairs (two user ids for one same people in different networks) provided by COSNET. They take advantage of the user information from [21] and [22] as the ground truth. The data is originally obtained by Perito *et al.* [22] by Google Profiles service in which it allows social network users to integrate their different accounts together. But instead of user pairs, what we need are group pairs. Here we introduce the basic idea to transform these user pairs into groups pairs. The previous research [23] shows that the one group usually contains some users with the large degree and many users with the small degree. Thus, inspired by the long tail distribution of the user degree, we classify the users of one group into two categories, i.e., the leader user with large degree whose degree is at the top of 20% of degree of all the users, while the common users are the 80% users whose degree is relatively small. Because we just know the user parts, our main idea is to calculate the proportion of the same ones of these two kinds of users in two groups.

TABLE I
Data Description

Network	Users	Relationships	Profile
Twitter	40,171,624	1,468,365,182	28,199
LiveJournal	3,017,286	87,037,567	2,101
Flickr	215,495	9,114,557	2,005
Last.fm	136,420	1,685,524	7,661
MySpace	854,498	6,489,736	9,993

The details are shown as followings. Suppose that g_i and g_j are two groups obtained by discovery algorithm from heterogeneous networks. Let $V_{g_i} = \mathrm{VL}_{g_i} \cup \mathrm{VC}_{g_i}$ and $V_{g_j} = \mathrm{VL}_{g_j} \cup \mathrm{VC}_{g_j}$ denote the users in these two groups, respectively, where VL are the leader users, and VC are the common users. Either $2(|\mathrm{VL}_{g_i} \cap \mathrm{VL}_{g_j}|)/(|\mathrm{VL}_{g_i}| + |\mathrm{VL}_{g_j}|) > 0.5$ or $2(|\mathrm{VC}_{g_i} \cap \mathrm{VC}_{g_j}|)/(|\mathrm{VC}_{g_i}| + |\mathrm{VC}g_j|) > 0.5$ meets the condition, then g_i and g_j are considered to be similar. These two functions mean that if the proportion of the same leaders or the same common users in these two groups is more than half, we would take these two groups as similar ones. In this case, not only the proportion of the two similar users in these two groups is guaranteed but also the considered degree of the users also contains the similarity of topology. At the same time, we know there are some flaws in this idea. We would present more insights on this problem in the discussion section later.

The model training dataset is randomly sampled, whose scale is $N/(1 + \mathrm{interval})$, where N is the total number of instances in the dataset, and interval is the sample interval. The rest ones are used for testing. By maximizing BDeu scores, we obtain the probabilistic graphs using training dataset. Based on these graphs and training data, we get the conditional probability of the two groups being similar. In order to reveal the effects of the two parameters equivalent sample size α and interval on our method, they are set as two segments, i.e., $\mathrm{interval} = \{1, 2, 3, 4, 5\}$ and $\alpha = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, and $\mathrm{interval} = \{5,\ 10,\ 15,\ 20,\ 25,\ 30,\ 35,\ 40,\ 45,\ 50\}$ and $\alpha = \{100,\ 200,\ 300,\ 400,\ 500,\ 600,\ 700,\ 800,\ 900,\ 1000,\ 1500,\ 2000\}$. The top-5 best probabilistic graphs are selected for the Bayesian model. All experiments are implemented on a 64-bit server with a Quad-Core 3.40GHz CPU and 16GB RAM.

In order to evaluate the validity of the detection results, we adopt three metrics including precision, recall, and $F1$-measure.

1) *Recall:* The proportion of the reported similar groups returned by the model.

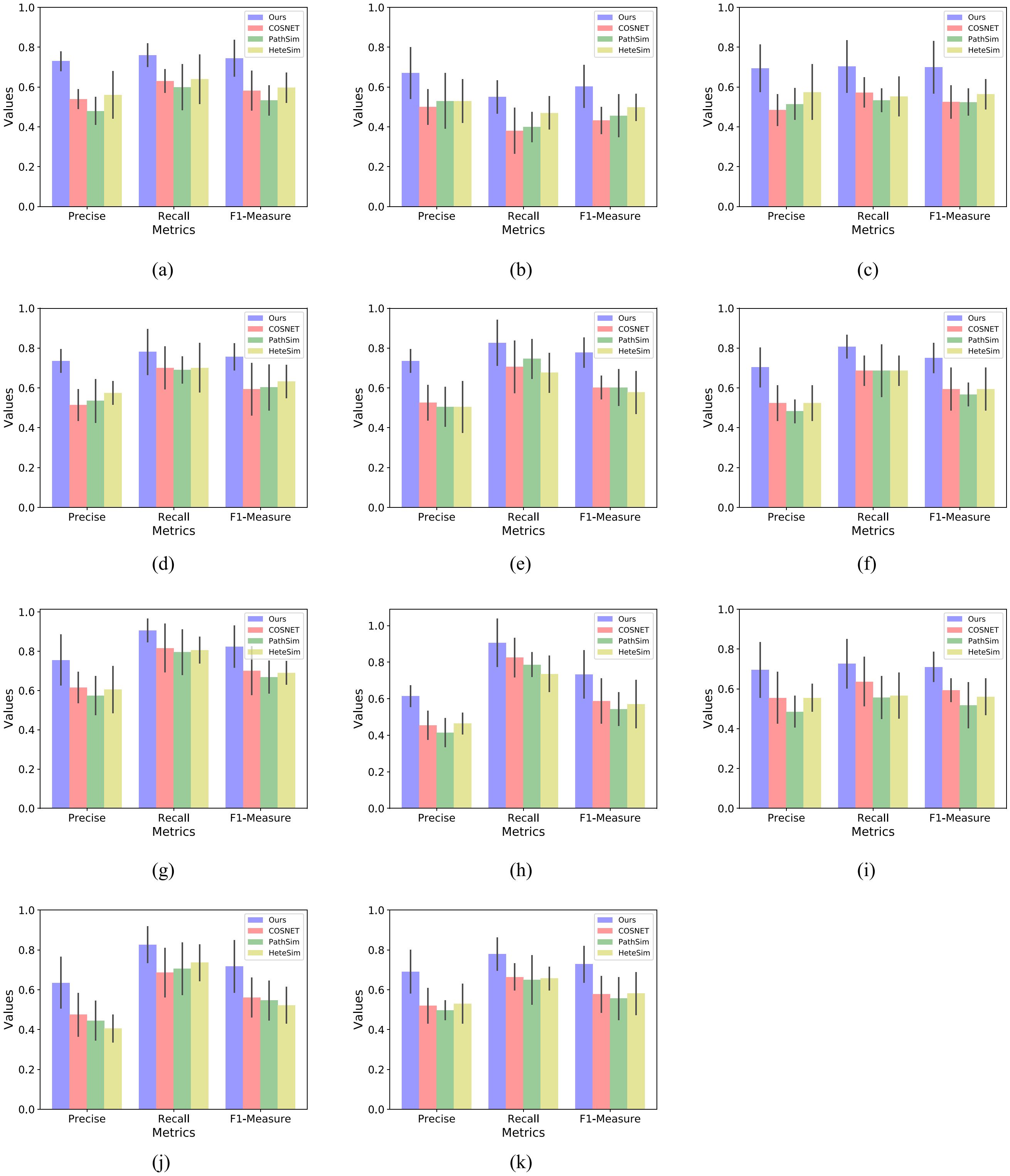


Fig. 9. Result comparison between our work and the existed related works. (a) Flickr-Last.fm. (b) Flickr-LiveJournal. (c) Flickr-MySpace. (d) Last.fm-LiveJournal. (e) Last.fm-MySpace. (f) LiveJournal-MySpace. (g) Twitter-Flickr. (h) Twitter-Last.fm. (i) Twitter-LiveJournal. (j) Twitter-MySpace. (k) Overall.

2) *Precision:* The proportion of the correct returned similar groups.
3) *F1-Measure:* The harmonic average of precision and recall.

B. Experimental Result

We introduce the details of our competitors and the comparison results. We also analyze the attribute contribution and the efficiency of our method. We also use a case study to illustrate

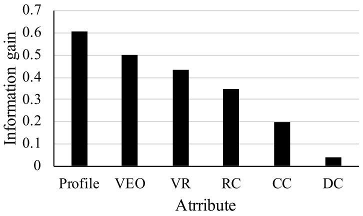


Fig. 10. Information gain of six adopted attributes in this paper.

the significance of our result to the performance improvement of other applications.

1) Competitors: Social media mining is a very hot topic in recent years. Several kinds of literature which try to discover the similar users in heterogeneous networks are highly related to this paper. We choose some of them as competitors and introduce the details of the implementation as follows.

COSNET [3]: They make use of the user profile and neighbors to calculate the similarity value of users in heterogeneous networks by employing a novel energy-based model.

PathSim [24]: They propose a meta-path-based method to use a meta path to represent the heterogeneous content of users in social networks, and infer the similarity degree of the users by computing the similarity value of their meta-paths.

HeteSim [25]: It is a similar method to PathSim. They take account in the heterogeneous meth-paths and obtain the user similarity based on these paths.

Because these previous works mainly focus on discovering similar users, which cannot solve the problem of matching groups directly. In order to use these methods, we take such strategies. First, we use MIRACLE [14] to detect similar groups in each network. Second, we take one group as one user whose profile is total profiles of all the users. Finally, we apply these methods on these coarsen networks, and find the similar groups replaced by detecting similar users.

2) Group Detection Results: We compare our method with the competitors in five different online social networks. The results adopt the average values with different parameters, which is shown in Fig. 9. The experimental result on each network pair and all the networks of the five different online social networks are presented.

From Fig. 9, we can see that the performances of our algorithm surpass the competitors in all three metrics. The improvements of precision, recall, and $F1$-measure are 20.8%–49.1%, 10.6%–44.7%, and 17.9%–47.6% over the compared algorithms. The overall improvements are 31.4%–44.3%, 17.3%–26.3%, and 25.9%–31.6% in precision, recall, and $F1$-measure, respectively. The better results verify the validity of our model. Although the useful information is very limited, all the methods obtain excellent performance, which verifies the validity of our idea.

3) Attribute Contribution Analysis: In order to show the contribution of the six attributes listed in Section III-B, we evaluate them by measuring the information gain with the respect to the class. The evaluation result is shown in Fig. 10.

Fig. 10 shows that the similarity of profile makes the largest significance for the result, whose information gain reaches 0.607. The similarity of vertex/edge overlapping follows the similarity of profile closely, whose values is 0.501. It means that the similar groups usually have similar users with similar profile and relationships across multiple online social networks. The similarity of degree centrality makes the least contribution, and mainly because the scale gap between these similar groups in different online social networks is relatively large. Based on the definition of similarity of degree centrality, this similarity would be affected by the large gap of group scales.

C. System Improvement

In order to verify the significance of the discovery results for other application, we adopt link prediction [26] as the case study, which is a very famous problem in the filed of social media mining research. We take Adamic–Adar method [27] as the baseline algorithm of link prediction. The Adamic–Adar index of u and v is calculated by $\sum_{w\in\Gamma(u)\cap\Gamma(v)} 1/\log \mid \Gamma(w) \mid$, where $\Gamma(u)$ denotes the set of neighbors of u.

Suppose that $gs = \{g_1, g_2, \ldots, g_k\}$ are the similar groups containing u and v across these different online social networks, we apply these similar information to improve the link prediction. We propose a simple function as follows:

$$f_i(u,v) = \begin{cases} 0, \text{ if } u \notin g_i \text{ or } v \notin g_i \\ 0.5, \text{ if } u \in g_i \text{ and } v \in g_i \text{ and } e(u,v) = 0 \\ 1, \text{ if } u \in g_i \text{ and } v \in g_i \text{ and } e(u,v) = 1 \end{cases} \tag{21}$$

where $e(u, v) = 1$ means that there is one edge between u and v, otherwise the opposite. The Adamic–Adar index based on similar groups is defined as

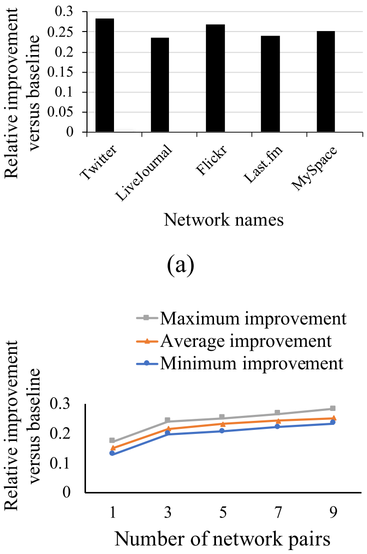
$$\left(1 + \sum\nolimits_{g_i\in gs} f_i(u,v)\big/|gs|\right) \sum\nolimits_{w\in\Gamma(u)\cap\Gamma(v)} 1/\log|\Gamma(w)|. \tag{22}$$

This function means that if u and v are in multiple same groups across online social networks, their Adamic–Adar index would be enhanced. The relative improvement of the enhanced algorithm using similar groups across multiple online social networks versus the baseline algorithm is presented in Fig. 11. The relative improvement is defined as $(F_e - F_b)/F_b$, where F_e and F_b are the $F1$-measures of results of enhanced algorithm and baseline algorithm, respectively.

Fig. 11(a) shows the relative improvement on different networks. When the enhanced algorithm predicts links on one network, it takes advantage of all the available similar groups to improve its performance. Although with small fluctuations, their improvements on different networks are around 0.25. We also show the scalability of the enhanced algorithm using similar groups ranging with the number of networks in Fig. 11(b). The curves include minimum improvement, average improvement, and maximum improvement. These improvements increase with the rising of the number of used networks. When the number of used network pairs reaches 5, the increasing trend tends to steady.

V. Discussion

In this section, we discuss the scalability of our algorithm.



(a)

(b)

Fig. 11. Relative improvement of link prediction using similar groups. (a) Relative improvement. (b) Improvement scalability.

A. Scalability

In this paper, we just take account in the user profile and relationship topology. If accessible, other information, such as check-in, travel information, public content and moment sharing, can also be used to improve the performance of our method. What is more, other group detection [28] and decision-making [29] algorithms could be applied to this problem.

Running time is always a challenge for the computation of big graphs. Group detection and topology similarity calculation occupy most of the running time. As one of the most popular methods to reduce the running time, parallel computing can also be used for our method. We can adopt multiple independent random walks to explore the graph structure. In this situation, the computation of this part can be parallelized as our previous work [14]. Because the computation of user similarity is independent, it is not hard to parallelize them. By splitting them into pieces and sending the pieces to computation nodes, the time cost will become much lower, which is beneficial to handle the large-scale social media. The running time on the different number of computer nodes is shown in Fig. 12. The running time is reduced significantly with the increasing computation resource. It reveals the potential ability to deal with the big data challenge.

VI. Related Work

Two lines of research can be used for supporting this paper.

Social Media Mining: With the sharply developing Internet, researchers from various kinds of fields attempt to exploit social media, although it is faced with the critical challenges of big data [30]. Among these, Yuan *et al*. [31] developed a data-driven method to detect function zones in a city based on the latent activity trajectories. Through mining the interaction pattern between e-commerce and social media, Zhao *et al*. [32] formulated a method which can deal with the *cold-start* problem in recommendation system.

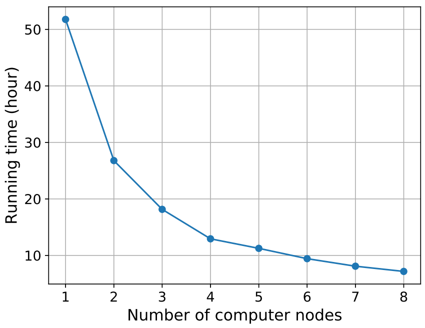


Fig. 12. Scalability of running time with different number of computer nodes.

What is more, the research about the heterogeneous content and relationships among networks is becoming a hot topic [33]. From the view of heterogeneous information networks, many researchers restart to study some classical problems, including user similarity, link prediction, and friend recommendation, etc. Pathsim [24] is a very popular idea to solve this problem, which calculates the user similarity value using semantic paths. Xiong *et al.* [34] proposed a method to get the similar user pairs by using the user-specified join path in the heterogeneous information network. Shi *et al.* [25] proposed a method named HeteSim, which could compute the closeness of the heterogeneous objects. COSNET [3] is a research paper focusing on connecting similar users in the heterogeneous networks. Both user profile and network topology are adopted in their work. They take advantage of the simple idea that the similar should share similar neighbors in different online social networks. Goga *et al.* [35] explored the study on how to choose reliable profiles for matching similar users across social networks. Shu *et al.* [36] summarized the related works about the user identity linkage across social networks.

Community Detection: Recent research identifies that group structure widely exists in various kinds of online social networks [4]. Community detection is one very classical and hot group research problem in the past years, which is related to this paper. This problem [28] has been thoroughly studied, and a community is defined as a group in which users have close relationships with each other [37]. The study of groups in online social networks can expose its function in the networks [38]. Newman [39] proposed a modularity optimization model to formulate this problem which has a deep effect on the research on community detection. By solving the optimization function for modularity, the community detection result is obtained. Pons and Latapy [40] tried to solve the community detection problem using random walks. They make use of random walks to measure the distance of users and merge them based on the distance. Liu *et al.* [14] proposed a parallel computing algorithm named *MIRACLE* based on multiple independent random walks, which can lower the running time of community detection significantly. Zhang *et al.* [41] started a research on the phase transition problem which exists in the community detection problem.

Atzmueller *et al.* [42] attempted to solve the description-oriented community detection problem by taking advantage of subgroup discovery. Zhang and Cao [43] developed a contact-burst-based clustering method which can discover transient communities by exploiting pairwise contact processes.

This paper is built on these works, especially in the fields of social media mining. Although there may be some similarities between this paper and the previous ones in social media mining, we focus on a very special and important filed, i.e., discovering similar groups across multiple online social networks. Compared with previous work, our proposed method has various kinds of advantages. First, the perspective of mapping different networks from groups benefits for mining the social data significantly, whose details have been presented in Section I. Second, our method is more easily parallelized. We employ random walks to explore the structure of networks, and the nature of the independence of random walks makes it very suitable for paralleling computing. We also detail the possible parallelism solution for our method in Section V-A with the valid experimental result for running time. This advantage makes our method able to handle the large-scale datasets well. Third, the Bayesian network model based on probabilistic graphs can not only obtain the inference result for the similarity of two groups in heterogeneous networks but also can get the causality relationship among these factors and inference result. All these advantages make our method dealing with this novel problem better than existing techniques.

VII. Conclusion

We propose and formulate an important problem, i.e., discovering similar groups across multiple social networks, which could improve the performance of other applications. In order to solve this problem, we propose an effective method: first, we use random walks to obtain group structure in each network, then extract the group similarity features from different angles, and infer whether two groups are similar or not based on the probabilistic graphical model, finally. We compare our algorithm with the state-of-the-art on five different online social networks. We also analyze the contribution of selected attributes and propose the possible solutions to solve the challenge of big data.

In the future work, we will explore more advanced models for the similar group detection problem. We will also expand the feature sources and polish the details of this paper to obtain better detection results. What is more, we will do our best to enhance our model and apply it to the mapping of groups in the virtual-network world and people in the real-life world.

References

[1] Y. Jiang and J. C. Jiang, "Diffusion in social networks: A multiagent perspective," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 2, pp. 198–213, Feb. 2015.

[2] L. Cui, J. Wu, D. Pi, P. Zhang, and P. Kennedy, "Dual implicit mining-based latent friend recommendation," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2017.2777889

[3] Y. Zhang, J. Tang, Z. Yang, J. Pei, and P. S. Yu, "COSNET: Connecting heterogeneous social networks with local and global consistency," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2015, pp. 1485–1494.

[4] J. H. Fowler and N. A. Christakis, "Cooperative behavior cascades in human social networks," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 12, pp. 5334–5338, 2010.

[5] C.-C. Wang, M.-Y. Day, and Y.-R. Lin, "Toward understanding the cliques of opinion spammers with social network analysis," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min. (ASONAM)*, San Francisco, CA, USA, 2016, pp. 1163–1169.

[6] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proc. ACM 21st Int. Conf. World Wide Web*, Lyon, France, 2012, pp. 191–200.

[7] Y. Shen and H. Jin, "Controllable information sharing for user accounts linkage across multiple online social networks," in *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manag.*, Shanghai, China, 2014, pp. 381–390.

[8] Z. Wang *et al.*, "Discovering and profiling overlapping communities in location-based social networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 4, pp. 499–509, Apr. 2014.

[9] J. Leng and P. Jiang, "Mining and matching relationships from interaction contexts in a social manufacturing paradigm," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 2, pp. 276–288, Feb. 2017.

[10] S. Agreste, P. De Meo, E. Ferrara, S. Piccolo, and A. Provetti, "Analysis of a heterogeneous social network of humans and cultural objects," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 4, pp. 559–570, Apr. 2015.

[11] M. R. Garey and D. S. Johnson, "The rectilinear Steiner tree problem is NP-complete," *SIAM J. Appl. Math.*, vol. 32, no. 4, pp. 826–834, 1977.

[12] R. Zafarani, L. Tang, and H. Liu, "User identification across social media," *ACM Trans. Knowl. Disc. Data*, vol. 10, no. 2, p. 16, 2015.

[13] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas.*, Melbourne, VIC, Australia, 2010, pp. 390–403.

[14] X. Liu, Y. Zhou, C. Hu, and X. Guan, "MIRACLE: A multiple independent random walks community parallel detection algorithm for big graphs," *J. Netw. Comput. Appl.*, vol. 70, pp. 89–101, Jul. 2016.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pageRank citation ranking: Bringing order to the Web," Stanford InfoLab, Stanford, CA, USA, Rep., 1999.

[16] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[17] D. Heckerman, C. Meek, and G. Cooper, "A Bayesian approach to causal discovery," in *Computation, Causation, and Discovery*, vol. 19. Cambridge, MA, USA: MIT Press, 1999, pp. 141–166.

[18] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press, 2000.

[19] D. Geiger and D. Heckerman, "A characterization of the Dirichlet distribution with application to learning Bayesian networks," in *Proc. 11th Conf. Uncertainty Artif. Intell.*, Montreal, QC, Canada, 1995, pp. 196–207.

[20] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Mach. Learn.*, vol. 20, no. 3, pp. 197–243, 1995.

[21] W. Chen, Z. Liu, X. Sun, and Y. Wang, "A game-theoretic framework to identify overlapping communities in social networks," *Data Min. Knowl. Disc.*, vol. 21, no. 2, pp. 224–240, 2010.

[22] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" in *Proc. Int. Symp. Privacy Enhancing Technol.*, Waterloo, ON, Canada, 2011, pp. 1–17.

[23] X. Liu, Y. Zhou, C. Hu, X. Guan, and J. Leng, "Detecting community structure for undirected big graphs based on random walks," in *Proc. ACM 23rd Int. Conf. World Wide Web*, Seoul, South Korea, 2014, pp. 1151–1156.

[24] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: Meta path-based top-K similarity search in heterogeneous information networks," *Proc. VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[25] C. Shi, X. Kong, Y. Huang, P. S. Yu, and B. Wu, "HeteSim: A general framework for relevance measure in heterogeneous networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, pp. 2479–2492, Oct. 2014.

[26] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Assoc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[27] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Soc. Netw.*, vol. 25, no. 3, pp. 211–230, 2003.

[28] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.

[29] G.-H. Tzeng and J.-J. Huang, *Multiple Attribute Decision Making: Methods and Applications*. Boca Raton, FL, USA: CRC Press, 2011.

[30] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[31] N. J. Yuan *et al.*, "Discovering urban functional zones using latent activity trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 712–725, Mar. 2015.

[32] W. X. Zhao *et al.*, "Connecting social media to E-commerce: Cold-start product recommendation using microblogging information," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1147–1159, May 2016.

[33] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, "A survey of heterogeneous information network analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 17–37, Jan. 2017.

[34] Y. Xiong, Y. Zhu, and P. S. Yu, "Top-k similarity join in heterogeneous information networks," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1710–1723, Jun. 2015.

[35] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi, "On the reliability of profile matching across large online social networks," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Sydney, NSW, Australia, 2015, pp. 1799–1808.

[36] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *ACM SIGKDD Explor. Newslett.*, vol. 18, no. 2, pp. 5–17, 2016.

[37] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Phys. Rep.*, vol. 424, nos. 4–5, pp. 175–308, 2006.

[38] S. Kelley, M. Goldberg, M. Magdon-Ismail, K. Mertsalov, and A. Wallace, "Defining and discovering communities in social networks," in *Handbook of Optimization in Complex Networks*. Boston, MA, USA: Springer, 2012, pp. 139–168.

[39] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 23, pp. 8577–8582, 2006.

[40] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Computer and Information Sciences—ISCIS 2005*. Heidelberg, Germany: Springer, 2005, pp. 284–293.

[41] P. Zhang, C. Moore, and M. Newman, "Community detection in networks with unequal groups," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 93, no. 1, 2016, Art. no. 012303.

[42] M. Atzmueller, S. Doerfel, and F. Mitzlaff, "Description-oriented community detection using exhaustive subgroup discovery," *Inf. Sci.*, vol. 329, pp. 965–984, Feb. 2016.

[43] X. Zhang and G. Cao, "Transient community detection and its application to data forwarding in delay tolerant networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2829–2843, Oct. 2017.

Xiaoming Liu is currently pursuing the Ph.D. degree with the School of Control Science and Technology, Xi'an Jiaotong University, Xi'an, China.

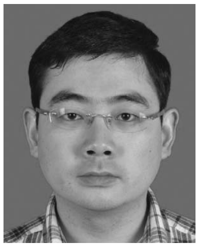
His current research interests include big graph mining, data analysis and mining, and network science and its applications.

Chao Shen (S'09–M'14) received the B.S. and Ph.D. degrees in control science and technology from Xi'an Jiaotong University, Xi'an, China, in 2009 and 2014, respectively.

He is currently an Assistant Professor with the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China. His current research interests include insider/intrusion detection, behavioral biometric, and measurement and experimental methodology.

Xiaohong Guan (S'89–M'93–SM'94–F'07) received the B.S. and Ph.D. degrees in control science and technology from Tsinghua University, Beijing, China, in 1989 and 1993, respectively.

He was a Senior Consulting Engineer with PGE, San Francisco, CA, USA, from 1993 to 1995. He was with the Division of Engineering and Applied Science, Harvard University, Cambridge, MA, USA, from 1999 to 2000. Since 1995, he has been with the Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, China, and has been the Cheung Kong Professor of Systems Engineering since 1999, and the Dean of the School of Electronic and Information Engineering since 2008. He has served as the Head of the Department of Automation from 2003 to 2008. His current research interests include allocation and scheduling of complex networked resources, network security, and sensor networks.



Yadong Zhou (M'10) received the B.S. and Ph.D. degrees in control science and technology from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2011, respectively.

He is an Assistant Professor with the Department of Automation, Xi'an Jiaotong University. He was a Post-Doctoral Researcher with the Chinese University of Hong Kong, Hong Kong, from 2014 to 2015. His current research interests include data analysis and mining, network science and its applications, and software defined networks.