Knowl Inf Syst (2014) 39:643–665
DOI 10.1007/s10115-013-0626-x

REGULAR PAPER

Remodeling the network for microgroup detection on microblog

Xiaobing Xiong · Gang Zhou · Xiang Niu · Yongzhong Huang · Ke Xu

Received: 6 March 2012 / Revised: 10 December 2012 / Accepted: 8 March 2013 /
Published online: 23 March 2013

Abstract In this paper, we focus on the problem of community detection on Sina weibo, the most popular microblogging system in China. By characterizing the structure and content of microgroup (community) on Sina weibo in detail, we observe that different from ordinary social networks, the degree assortativity coefficients are negative on most microgroups. In addition, we find that users from the same microgroup tend to share some common attributes (e.g., followers, tags) and interests extracted from their published posts. Inspired by these new findings, we propose a united method to remodel the network for microgroup detection while maintaining the information of link structure and user content. Firstly, the link direction is concerned by assigning greater weight values to more surprising links, while the content similarity is measured by the Jaccard coefficient of common features and interest similarity based on Latent Dirichlet Allocation model. Then, both link direction and content similarity between two users are uniformly converted to the edge weight of a new remodeled network, which is undirected and weighted. Finally, multiple frequently used community detection algorithms that support weighted networks could be employed. Extensive experiments on real-world social networks show that both link structure and user content play almost equally important roles in microgroup detection on Sina weibo. Our method outperforms the traditional methods with average accuracy improvement up to 39 %, and the number of unrecognized users decreased by about 75 %.

Keywords Microblogging · Community detection · Link structure · User content · United method

X. Xiong (✉) · G. Zhou · Y. Huang
State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China
e-mail: bingxiaoxiong@gmail.com

X. Xiong · X. Niu · K. Xu (✉)
State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
e-mail: kexu@nlsde.buaa.edu.cn

1 Introduction

In recent years, the microblogging system has emerged as a novel social media platform and provides us a new means of communication, with which users can broadcast brief updates about any thing happening in their daily life or work, such as what they are doing, watching, or thinking about. The most famous microblogging system in the world is Twitter, while the most famous one in China is Sina weibo, which began in August 2009 and has gained more than 200 million users as of December 2011.

Sina weibo introduced a new feature called *Microgroup* in November 2010. A microgroup is usually a group of users who have close connections or share similar interests. The number of microgroups has been growing rapidly, and there are more than 3 million microgroups which contain more than 80 million users as of December 2011. The emergence of microgroup provides an unprecedented opportunities to study user classification (i.e., community detection). Community detection within real-world networks, such as online social networks (OSNs) and biological networks, is a problem of considerable practical interest and has attracted a great deal of attention [3,8,14,17,32,33,35,41,42,46,50]. Thus, we believe that community detection on microblogging systems (e.g., Twitter, Sina weibo) is also very worthy of investigation in depth.

In this paper, Sina weibo is chosen as our experimental platform for community detection on microblogging systems. As a community (i.e., microgroup) is a group of users not only with close connections, but also with similar interests on Sina weibo, we focus on determining community memberships by using link relationship and user content simultaneously, and a new united method stressing both factors is presented to group individuals with higher accuracy on microblogging systems.

The remainder of this paper is organized as follows. In Sect. 2, we give an overview of the related work. Section 3 describes the details of our data set. In Sect. 4, we characterize the sampled microgroups in detail mainly from aspects of network structure and user content, and conclude some useful findings for microgroup detection. A united method to remodel the network for microgroup detection on Sina weibo is proposed in Sect. 5. Section 6 introduces some widely used methods to measure the similarity of different partitions. In Sect. 7, we apply the new method on several real-world networks and compare their outcomes against some traditional algorithms. Finally, in Sect. 8, we conclude this paper and discusses future research briefly.

2 Related work

Most previous community detection approaches are mainly based on structural features (e.g., links), and a community is usually defined as a group of vertices such that there is a higher density of edges within groups than between them [6,13]. Then, an objective function named modularity degree is often used to capture the above intuition of a community [32]. As the objective is typically NP-hard to optimize, many algorithms, including spectral partitioning [34], hierarchical clustering, heuristics, and approximation solutions [23], have been extensively studied. However, most of the traditional algorithms based on modularity optimization have some drawbacks, such as needing user-specified number of communities, considering no link direction or weight, concerning no overlap among communities, and being not suitable for large-scale networks. Thus, in recent years, many novel researches have yet been done on community detection. Gregory et al. [18] proposed an algorithm for finding overlapping community structure in very large networks based on the label propagation technique

of Raghavan et al. [36]. Lai et al. [27] address the problem of finding communities on directed networks by using PageRank random walk induced network embedding to transform a directed network into an undirected one, where the information on edge directions is effectively incorporated into the edges weights. Yang et al. [47] presented a probabilistic model for community detection in directed networks that aims to model both incoming links and outgoing links simultaneously and differentially. Duan et al. [10] proposed a two-step approach to discover the community structure of a weighted and directed graph in one time-slice. Stanoev et al. [38] presented a novel algorithm for community detection that combines network structure with the processes that support creation and/or evolution of communities.

Since microblogging systems like Twitter and Sina weibo do more in sharing and spreading information than maintaining friendship, many users choose to follow others for being interested in their posts. Hence, the link induced by "follow" behavior implies more on interest than friendship. Except for link structure, microblogging systems also open member's content information (e.g., user attributes, published posts, discussed topics), which is very useful for extracting user's interest. Then, if only user's interest is considered, many clustering algorithms can be employed for community detection on microblogging networks, including partitioning clustering algorithm [2,20], hierarchical clustering algorithm [43,51], evolutionary clustering algorithm [21], density-based clustering algorithm [1,12], model-based clustering algorithm [5], graph-based clustering algorithm [2], and synthetic algorithm [22].

However, several previous works have found that neither link structure nor user content is sufficient to determine the community memberships while combing link with content usually achieves better performance [7,16]. For example, Erosheva et al. combined LDA with LDA-Link for network analysis [39]. Yang et al. [48] proposed a discriminative model for combining the link and content analysis for community detection from networked data, such as paper citation networks. Other approaches that exploit topic models for community detection include [9] and [19]. However, these previous researches are very similar to text classification and act only on the networks with nodes denoting text pages (e.g., blog pages, wikipedia pages, and published papers), but not on the social networks with nodes denoting human individuals. When detecting community for text networks, we should consider only the "citing" relationship and the content similarity between texts. While, in order to classify individuals on microblogging network, we should consider both the "following" relationship and the interest similarity between individuals. As a complex human feature, individual interest should be measured considering many factors like published posts, discussed topics, labeled tags, etc.

Basically, remodeling the microblogging network with a weighted network is the kernel of our work, then it is very significant to measure the edge weight between two users. Predicting the strength of social ties is a well-studied problem in the field of social network analysis [52], such as Xiang et al. [44] developed a latent variable model to estimate relationship strength from interaction activity and user similarity. However, most of those studies have only tried to measure the strength of existing links in the networks but did not care about the edge weight between users without connection, which is also helpful for our task of community detection on microblogging networks. Furthermore, none of those previous methods have measured the edge weight from the perspective of community detection, in which a higher strength should be assigned to the social tie between users from the same community than that from different communities. Furthermore, the interaction activity and user profile employed in those previous papers were selected experimentally, but not by extensive statistical analysis on real-world social networks, and this may result in adverse impact on the accuracy of community detection on microblogging networks. Thus, it is a challenging but rewarding

task for us to cluster homogeneous users with compact connection and similar interest into communities on microblogging networks.

An earlier version of this paper [45] was presented at the 7th International Conference on Advanced Data Mining and Applications.

3 Data set

On Sina weibo, microgroup was introduced in November 2010 and attracted more than 80 million users as of December 2011. As most microgroups are formed by users with similar interest, therefore, in order to analyze the explicit and internal characteristics of microgroups, we selected 34 microgroups, covering almost all kinds of microgroups on Sina weibo, for analysis. We crawled the link structure and content information of these microgroups in a short time interval from March 6 to March 21, 2011. For each member of the 34 microgroups, we crawled their profile (including user ID, name, location, gender, verified flag), social network (i.e., both followers and followings lists), and posted content (including the tags and topics adopted in the posts). In total, we crawled about 200 thousand users from the 34 microgroups.

Each microgroup represents a community of Sina weibo users and can be characterized as a directed graph, in which a vertex represents a member, and an edge indicates a "following" relationship. Specifically, a directed edge from A to B indicates that user A follows B. In other words, A is a follower of B, and B is a following of A. Note that a directed graph for a microgroup only includes "following" relationship between its members while ignoring edges from or to users outside the microgroup.

4 Characterizing microgroup

To start with, we explore ways to characterize microgoups and see whether some observations while analyzing the characteristics of the microgroups can lead us to more efficient way of microgroup detection.

We begin our analysis of microgroup with the following questions: (i) Whether the users linked by a "following" relationship share similar interests, or attributes? (ii) What are the most important factors driving users to join in a particular microgroup? and (iii) Do members of a microgroup post similar content? To answer these questions, we first summarize the basic information of the sampled microgroups and characterize the groups using both features of network structure and content.

4.1 Basic analysis

We summarize the basic network properties of the sampled microgroups, and the results are shown in Table 1, from which we can know that those networks are very sparse with low densities (with an average density of 0.0077), and the mean degrees of most microgroups are much lower than those of traditional social networks. Furthermore, about one in five members are isolated and have no link relationship with others, which indicates those members join microgroups just for sharing information, but not making friends. Hence, isolated users cannot be classified only from the view of link structure; however, their attributes may help to explain why isolated users decide to join the microgroup, and they may be similar to others on some attributes or interests.

Table 1 Basic statistical characterizations of the sampled microgroup networks

ID	Node#	dEdge#	BiE (%)	MD	IsoN (%)	Density
245808	6456	14945	36.63	2.3	19.45	0.0004
103797	1297	2379	39.37	1.8	29.31	0.0014
106580	7097	20581	28.6	2.9	18.75	0.0004
110109	2601	28404	56.69	10.9	14.01	0.0042
121002	992	1024	44.23	1	33.27	0.0010
124954	6110	44631	52.48	7.3	16.93	0.0012
132605	661	1266	32.98	1.9	21.32	0.0029
166330	2494	24343	60.87	9.8	14.09	0.0039
175812	1752	4245	58.99	2.4	15.28	0.0014
200627	3002	30263	46.87	10.1	10.87	0.0034
106940	724	1221	40.02	1.7	31.61	0.0023
138755	325	5495	29.05	16.9	10.86	0.0522
183101	389	2978	56.49	7.6	22.04	0.0363
223539	223	229	37.13	1	26.32	0.0046
227065	3471	38329	52.71	11	15.45	0.0032
240466	2134	47574	59.93	22.3	12.64	0.0105
248799	2301	53530	53.76	23.3	6.31	0.0101
249397	1735	4108	48.68	2.4	30.27	0.0014
268147	570	3201	57.53	5.6	13.05	0.0047
272059	1080	1579	49.24	1.5	20.31	0.0014
281493	6526	51650	50.93	7.9	11.58	0.0012
287686	875	8155	45.08	9.3	9.38	0.0107
101126	334	332	24.81	1	24.84	0.0030
209826	1436	3433	42.15	2.4	20.92	0.0017
156447	253	1362	41.43	5.4	12.44	0.0214
172122	567	1739	58.38	3.1	16.83	0.0054
235497	122	95	15.85	0.8	39.05	0.0064
250802	124	308	47.37	2.5	24.47	0.0202
257229	98	57	18.75	0.6	45.46	0.0060
271053	405	2167	44.95	5.4	16.32	0.0132
288096	920	2008	27.65	2.2	22.52	0.0024
299815	264	1088	69.73	4.1	16.46	0.0157
168175	2828	21618	39.56	7.6	13.73	0.0027
228209	930	3617	54.9	3.9	18.87	0.0042
Avg.	1797	12587	44.82	5.9	19.85	0.0077

ID microgroup ID; *Node#* number of nodes; *dEdge#* number of directed edges; *BiE* (%) ratio of bi-way edges; *MD* mean degree; *IsoN* (%) ratio of isolated nodes; *Density* network density

The ratio of bi-edges is often used to measure the reciprocity of a social network, and many previous studies have reported high level of reciprocity on some social networks: 68 % of user pairs with any link between them are connected bi-way on Flickr [4] and 84 % on Yahoo!360 [25]. However, [26] observed a low level of reciprocity on Twitter: 77.9 % of

user pairs are one-way, and only 22.1 % have reciprocal relationships between them. The low reciprocity on Twitter might be accounted for the billions of one-direction links from normal users to celebrities and other power users on Twitter. As for the sampled microgroups, the level of reciprocity is moderate: 44.8 % of user pairs are bi-way, which is larger than Twitter but lower than Flickr. As similar with Twitter, most celebrities and mass media are followed by a large number of users but do not follow them back, what's more, few celebrities and media would like to join a microgroup on Sina weibo.

4.2 Assortativity coefficient

Similarity breeding connection is a principle that structures many kinds of network ties, including friendship, reference, support, coauthor. We call the principle as homophily, an important criterion to quantify the tendency for users to be friends with others who have similar characteristics, which can be measured by assortativity coefficient. Here, in order to analyze the similarity degree among users from the same microgroup, we study the assortativity coefficients by different user attributes, including *degree*, *gender*, *location*, *VFlag*, *posts count*, *reposts count*, and *comments count*.

On undirected networks, degree assortativity coefficient is often calculated as the Pearson's correlation coefficient of the degrees at either ends of the edges. Then, Newman proposed a method to calculate the degree assortativity coefficient for directed network [31], defined as:

$$r = \frac{\sum_i j_i k_i - M^{-1} \sum_i j_i \sum_i k_i}{\sqrt{\left[\sum_i j_i^2 - M^{-1}\left(\sum_i j_i\right)^2\right]\left[\sum_i k_i^2 - M^{-1}\left(\sum_i k_i\right)^2\right]}} \tag{1}$$

where j_i and k_i are the excess in-degree and out-degree of the vertices that the ith edge leads into and out of, respectively, and M is the total number of edges. The value of r lies in the range of $-1 \leq r \leq 1$, with $r = 1$ indicating perfect assortativity; $r = 0$ indicating no assortativity; and $r = -1$ indicating perfect disassortativity (i.e., perfect negative correlation). Here, Eq. 1 is used to calculate the degree assortativity coefficient of microgroup.

Furthermore, when calculating the assortativity coefficient by other continuous attribute like *posts count*, *reposts count*, and *comments count*, we measure by calculating the standard Pearson's correlation coefficient [31]; thus

$$r = \frac{\sum_{xy} xy \left(e_{xy} - a_x b_y\right)}{\sigma_a \sigma_b} \tag{2}$$

and

$$\sum_{xy} e_{xy} = 1, \sum_y e_{xy} = a_x, \sum_x e_{xy} = b_y$$

where e_{xy} is defined as the fraction of all edges in the network that join together vertices with values of x and y for the continuous variable attribute like *posts count*. a_x and b_y are, respectively, the fraction of edges that start and end at vertices with values of x and y. σ_a and σ_b are the standard deviations of the distributions a_x and b_y. The value of r has the same interpretation with degree assortativity coefficient in Eq. 1.

However, when computing assortativity coefficient by discrete or enumerative attribute like *gender* and *location*, a general measure of scalar assortativity relative to a categorical

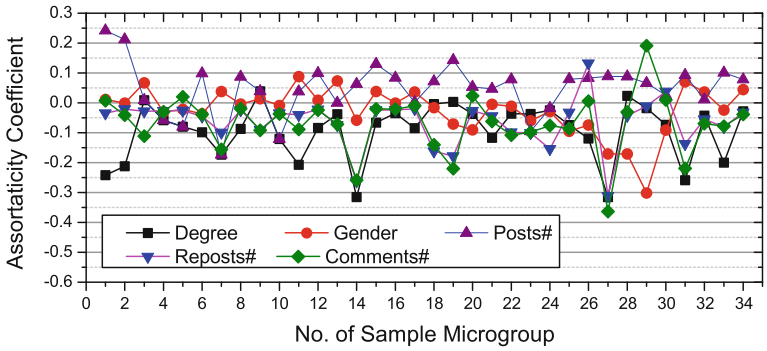

Fig. 1 Assortativity coefficients by different user attributes (*Degree*, *Gender*, *Posts#*, *Reposts#*, *Comments#*) on each microgroup

variable is given by

$$r = \frac{tr(e) - \|e^2\|}{1 - \|e^2\|} \tag{3}$$

where $e = E/\|E\|$ is the normalized mixing matrix, the elements E_{ij} give the number of edges in the network that connect from a node of type i (e.g., users from "Beijing") to a node of type j (e.g., users from "Shanghai").

The results of assortativity coefficients based on some user attributes are described in Fig. 1, from which we can know that although many ordinary social networks tend to be positively assortative with respect to *degree*, for most of the sampled microgroups, the degree assortativities (denoted by ■ in Fig. 1) are negative, or weakly positive with low values, which implies that most users with few followers tend to follow others with many followers, like celebrities. Besides, the two users with "follow" relationship are not so similar on the attribute of *degree* in most cases.

To some extent, the number of posts can be used to measure a user's activity on Sina weibo. From our observations, we find that many assortativity coefficients by *posts#* (denoted by ▲ in Fig. 1) are positive, which is distinct from degree assortativity and indicates that many users tend to follow others with similar number of published posts. In addition, the number of reposts and comments are important measurements for user's popularity on Sina weibo, and assortativity coefficients by *reposts#* (denoted by ▼ in Fig. 1) imply a similar conclusion with degree assortativity. That is, most ordinary users tend to follow others with high popularity. Assortativity coefficients by *comments#* (denoted by ♦ in Fig. 1) show a consistent result with *reposts#*.

Assortativity coefficients by *gender* (denoted by • in Fig. 1) are very close to 0 on most of the sampled microgroups, which reveals that there is no obvious tendency for user to follow others with the opposite or the same gender. As most members on microgroups are non-verified, and some microgroups like collegiate groups are localized in the same city, assortativity coefficients based on *VFlag* and *location* will be ignored in this paper.

By analyzing the assortativity coefficients based on different user attributes, we find that the two ends with link relationship are different from each other on many attributes like degree and popularity, which differs significantly from other ordinary social networks. For this reason, when classifying users to different microgroups, link structure is not satisfactory in determining accurately the community memberships.

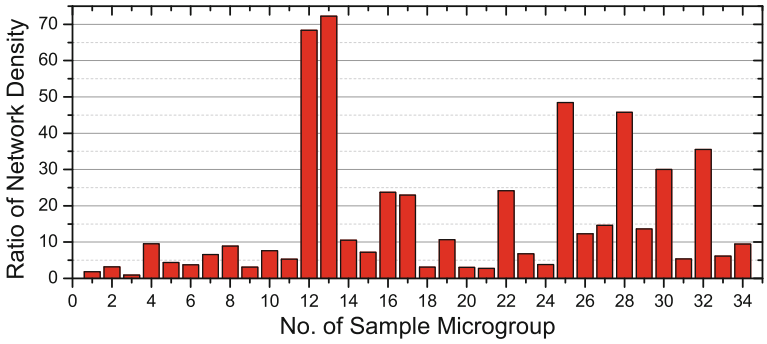

Fig. 2 Densities of the sampled microgroups compared to random case

4.3 Density difference

In common definition, a community is a subset of users within which the network connections are dense, but between which they are sparser. So the densities of the sampled microgroups should be higher than the random sampling groups. In this paper, the density of a directed network $G(n, m)$ is defined as $d = \frac{m}{n \times (n-1)}$, and m, n are, respectively, the total number of directed edges and nodes in the network.

In Fig. 2, we show the densities of the sampled microgroups, which are presented by the ratios compared to the random case. From Fig. 2, we know that the densities of the sampled microgroups vary a lot, and many have much more compact structure than random sampling groups, like microgroup 12, 13, 25, 28, and 32. However, there are also some microgroups with density close to the random case, such as microgroup 3, 1, 21, 20, 18, and 9; besides, the density of microgroup 3 is less than the random case. The existence of these microgroups with low densities breaks the traditional definition of community, which focuses only on the compactness of the link structure. Thus, we further enhance our inference that it is not sufficient to identify microgroups by only considering the link structure on Sina weibo.

4.4 Attribute similarity

In many social networks, attribute similarity is a basic principle for users to gather together in the same community. In order to dig out the distinctive characteristics of members from the same microgroup, we analyze the average similarity among users based on their common *followers*, *followings*, *tags*, and *topics*, respectively, on each microgroup, then compare the observation results with a series of random sampling groups. By comparison, we try to reveal what characteristics are remarkable for us to label users from the same microgroup, which can help to identify communities on Sina weibo.

In this paper, Jaccard coefficient, a commonly used similarity metric in information retrieval [29], is used to measure the attribute similarity among users, i.e., the probability that both x and y have common feature f. If the "feature" here is taken as *followers*, *followings*, *tags*, or *topics*, the similarity between user x and y on each feature can be calculated as follows:

$$s_f(x, y) = \frac{\left|\Gamma_f(x) \cap \Gamma_f(y)\right|}{\left|\Gamma_f(x) \cup \Gamma_f(y)\right|} \tag{4}$$

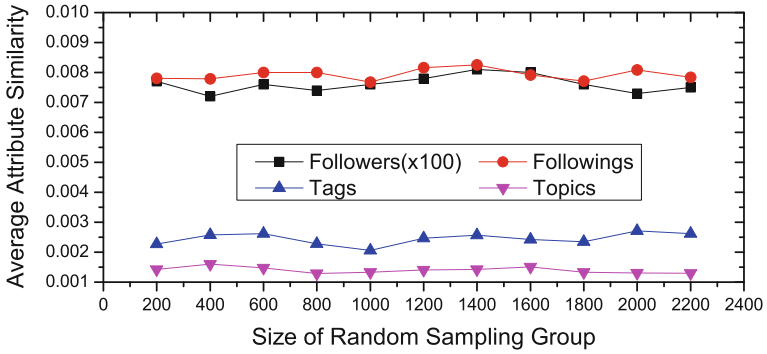

Fig. 3 Average attribute similarity measured by Jaccard coefficient on random sampling groups. The considered attributes are *Followers*, *Followings*, *Tags*, and *Topics*

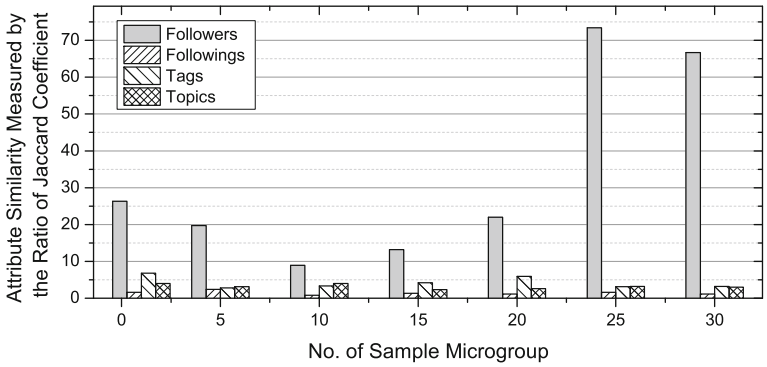

Fig. 4 Attribute similarity of the sampled microgroups compared to random case

where $\Gamma_f(x)$ is the set of feature f for user x, such as the set of followers of user x, and $|\Gamma|$ is the number of elements in Γ. Then, on each microgroup, the average similarity among users by different features is as follows:

$$avg\_s_f(G) = \frac{1}{|\Gamma(G)| \times (|\Gamma(G)| - 1)} \sum_{x,y \in \Gamma(G)} s_f(x, y) \tag{5}$$

where $\Gamma(G)$ is the set of users on microgroup G.

We have mentioned that a microgroup should be a group of users with similar interests measured by users' feature and content. In order to estimate what features are more prominent, we construct a series of random sampling groups with different number of users from all crawled users. The average attribute similarities in random sampling groups are shown in Fig. 3, from which we see that, with the increase in the size of random sampling group, the average attribute similarity measured by Jaccard coefficient is very stable on all considered features. Thus, our approach of random sampling is nearly unbiased, and the result of random case can be seen as a baseline.

The results of average similarity by different features on some sampled microgroups are shown in Fig. 4, which are expressed by the ratios compared to the random sampling groups described in Fig. 3. We call the ratio as significance degree in this paper. From Fig. 4, we find that the feature of *followers* is the most significant and the average ratio

is much higher (average: 22.6) than the other three features, in which *followings* is the least prominent and most ratios are close to one (average: 1.4), i.e., the average similarity by *followings* on the sampled microgroups is very close to random sampling group. For this reason, we conclude that users with more common *followers* are more likely to be similar and from the same microgroup, but this is not the case for *followings*. The non-significance of the feature of *followings* may be induced by the truth that many users choose to follow some common celebrities simultaneously; however, the activity of celebrity following expresses little about individual interest. Intuitively, the features of *tags* and *topics* are obvious indicators of individual interest, but our results in Fig. 4 show that the two features are not so significant as expected, and with the average ratios of Jaccard coefficients about 3.8 and 3.6, respectively. Thus, we conclude that the feature of *followers* is the most significant for microgroup detection, then followed by *tags* and *topics*, but the feature of *followings* is nearly indistinctive, and will be ignored when identifying microgroup.

4.5 Interest similarity

To identify individual topical interests of a particular user, we apply the well-known Latent Dirichlet Allocation (LDA) model on collections of posts from users. LDA model uses a "bag of words" assumption, which treats each document as a vector of word counts. Based on this assumption, each document is represented as a random mixture over latent topics, denoted as $P(z)$, where each topic is characterized by a distribution over a number of words, denoted as $P(w|z)$. Therefore, each word within a document can be calculated as follows:

$$P(w_i) = \sum_{j=1}^{T} p(w_i|z_i = j)p(z_i = j) \tag{6}$$

LDA is the first complete topic model, and the document's generative process can be graphically represented using commonly used plate notation in Fig. 5. In this figure, shaded and unshaded plates indicate observed and latent variables, respectively. An arrow corresponds to a conditional dependency between two variables and boxes indicate repeated sampling with the number of repetitions given by the variable in the bottom of the corresponding box. Formally, each of a collection of D document is associated with a multinominal distribution over T topics, which is denoted as θ. Each topic is associated with a multinominal distribution over words, denoted as ϕ. θ and ϕ have Dirichlet prior with hyper-parameters α and β, respectively. For each word in one document d, a topic z is sampled from the multinominal distribution θ associated with the document, and a word w from the multinominal distribution ϕ associated with topic z is sampled consequently. Above generative process is repeated N_d times (N_d is the total number of words in document d) to form document d. Given the parameters α and β, the joint distribution of a set of N topics $\mathbf{z}$ and a set of N words $\mathbf{w}$ is given by:

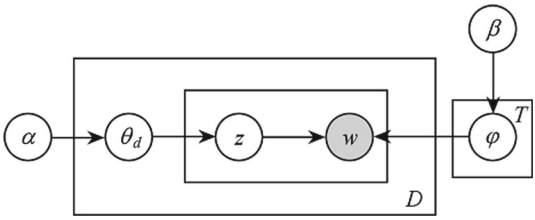

Fig. 5 Bayesian network of LDA model

$$p(\theta, z|\alpha, \beta) = p(w|z, \beta)p(z|\alpha) \int p(z|\theta)p(\theta|\alpha)\mathrm{d}\theta \int p(w|z, \phi)p(\phi|\beta)\mathrm{d}\phi \tag{7}$$

LDA model has two parameters to be inferred from the data, i.e., document-topic distributions θ and the T topic-word distribution ϕ. The words in the document can be observed, while the topic mixture is latent. According to the generative process and known words, above two parameters can be estimated. In this study, Gibbs sampling is applied for model parameter estimation. To extract user's interest based on LDA model, document should naturally correspond to published posts. However, since the goal is to understand what each user is interested in rather than what each published post is about, we aggregate all the posts published by individual user into a big document. Thus, each document essentially corresponds to an individual. The result of LDA model can be mainly represented by three matrices:

- DT, a $D \times T$ matrix that contains the document-topic distributions, where D is the number of documents and T is the number of topics. DT_{ij} captures the probability that document d_i has been assigned to topic t_j, i.e., the probability that individual s_i is interested in topic t_j.
- WT, a $W \times T$ matrix that contains the word-topic distributions, where W is the number of unique words used in all documents and T is the number of topics. WT_{ij} captures the probability that word w_i has been assigned to topic t_j.
- Z, a $1 \times N$ vector, where N is the total number of all words in the bag of documents. Z_i is topic assignment for word w_i.

Among above three matrices in the result of LDA model, the matrix DT is the most important for us to measure the interest divergence between users. As each row of DT is basically the probability distribution of individual interest over the T topics, the interest divergence between user s_i and s_j can be calculated as [11]:

$$dist(i, j) = \sqrt{2 \times D_{JS}(i, j)} \tag{8}$$

where the value of $dist(i, j)$ lies in the range $0 \leq r \leq 1$. $D_{JS}(i, j)$ is the Jensen–Shannon divergence between the two probability distributions DT_i and DT_j, which is defined as:

$$D_{JS}(i, j) = \frac{1}{2}(D_{KL}(DT_{i.}||R) + D_{KL}(DT_{j.}||R)) \tag{9}$$

where R is the average of the two probability distributions, i.e., $R = \frac{1}{2}(DT_{i.} + DT_{j.})$. D_{KL} in above equation is the Kullback–Leibler divergence [40] used to define the divergence from distribution Q to P.

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{10}$$

Therefore, from the perspective of published posts, the average divergence of interest among users from the same microgroup can be calculated as:

$$avg\_d(G) = \frac{1}{|\Gamma(G)| \times (|\Gamma(G)| - 1)} \sum_{i,j \in \Gamma(G)} dist(i, j) \tag{11}$$

where $\Gamma(G)$ is the set of users on microgroup G.

We calculate the average interest divergence for each sampling group based on LDA model, and the results are shown in Fig. 6, from which we know that with the varying of the size of sampling group and the assigned number of latent topics, the average interest divergence based on LDA model is almost consistent nearby 0.71. For this reason, we say

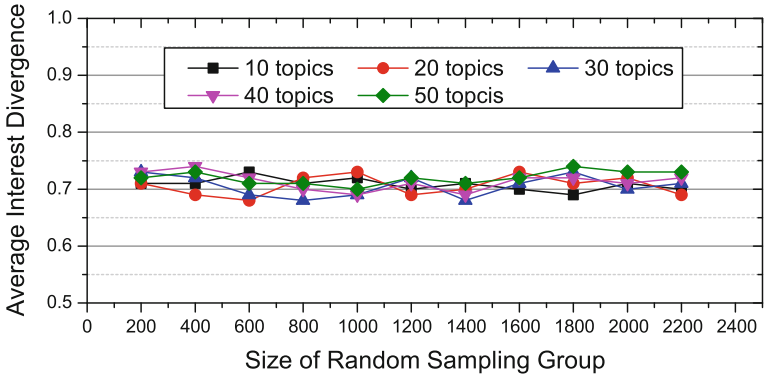

Fig. 6 Average interest divergence on random groups

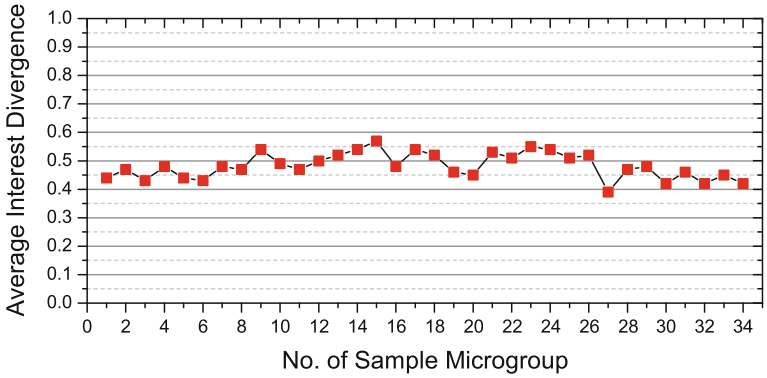

Fig. 7 Average interest divergence on the sampled microgroups

that average interest divergence on random sampling groups is unbiased and can be seen as a baseline. As the interest divergence is insensitive to the assigned number of latent topics, we set it to 30 for later discussion.

Figure 7 illustrates the results of the average interest divergence on the sampled microgroups, which are all around 0.5 and distinctly lower than random case (i.e., 0.71). That is, from the perspective of published posts, the users from the same microgroup may have more common topics, i.e., share more similar interests than random case. Hence, individual interest extracted from published posts can be used as an important factor for microgroup detection.

5 United method for microgroup detection

Microblogging users construct an unweighted and directed network, and the problem of community detection on directed networks has been well studied. However, as aforementioned, neither link structure nor user content is satisfactory in identifying the community memberships on Sina weibo: the link relationship is usually sparse on microblogging networks, and the two ends with link relationship may be different on some user attributes and interests. In addition, the irrelevant user attributes may mislead the result of microgroup detection.

Hence, in order to solve the problem of community detection on Sina weibo with a higher accuracy and performance, we propose a united method without losing the information of link structure and user content. Here, user content mainly includes published posts and user attributes like followers list and tags list. Then, there are two "links" between users, one is the explicit "following" relationship, and the other is the implicit attribute similarity between the two ends. With our new method, we uniformly convert the link structure and content similarity to the edge weight of a new remodeled network, then many well-known community detection algorithms that support weighted networks would be employed.

Consider two nodes i and j on Sina weibo, the edge weight between the two nodes can be calculated as follows:

$$W_{ij} = \alpha L'_{ij} + \beta S'_{ij} \tag{12}$$

where α, β $(\alpha + \beta = 1)$ are, respectively, the weight value of link structure and content similarity in microgroup detection. L'_{ij} is the normalized edge weight converted from the information of link direction, while S'_{ij} is the normalized interest similarity extracted from user attributes and published posts between the two users.

In our united approach, instead of simply ignoring directional information, we use the method proposed by Kim et al. [24] to convert the information of link direction to the weight of a new undirected link. The key idea is to give higher weight to the more *surprising* link, and the *surprising* degree is measured by the probability of the link. Let us consider a link directing from node i to j, and the probability of this link, when the links are assigned randomly while keeping the degree of each node, is

$$p_{ij} = \frac{k_i^{\text{out}} k_j^{\text{in}}/2m}{k_i^{\text{out}} k_j^{\text{in}}/2m + k_j^{\text{out}} k_i^{\text{in}}/2m} \tag{13}$$

where $k_i^{\text{out}} = \sum_j A_{ij}$ and $k_j^{\text{in}} = \sum_i A_{ij}$ are, respectively, the outgoing and incoming degree of node i and j, where $A_{ij} = 1$ if there is a link from i to j, and 0 otherwise, and m is the total number of links calculated as $m = \sum_i \sum_j A_{ij}$.

As smaller p_{ij} indicates stronger relatedness for the direction from node i to j, then the weight of the link between node i and j is defined as

$$L_{ij} = A_{ij}(1 - p_{ij}) + A_{ji}(1 - p_{ji}) \tag{14}$$

$\{L_{ij}\}$ is an undirected and weighted network transferred from the original directed and unweighted network $\{A_{ij}\}$.

When measuring the content similarity between two users, we consider two aspects: one is the attribute similarity, and the other is the interest similarity. In our work, attribute similarity is measured by the Jaccard coefficient of user features stated above, from which we know that the average similarity on the feature of *followers* is the most prominent, and followed by *tags* and *topics*. However, the average similarity on the feature of *followings* is nearly the same as random case and will be ignored when measuring the attribute similarity between two users. In addition, the interest similarity will be calculated based on LDA model. Thus, we define the content similarity between user i and j as

$$S_{ij} = \theta_1 \left(\eta_1 s_{ij}^{\text{fol}} + \eta_2 s_{ij}^{\text{tag}} + \eta_3 s_{ij}^{\text{top}} \right) + \theta_2 (1 - dist(i, j)) \tag{15}$$

where θ_1 and θ_2 are, respectively, the weight value of attribute similarity and interest similarity, and will be assigned by empirically analyzing. s_{ij}^{fol}, s_{ij}^{tag}, and s_{ij}^{top} are, respectively,

the normalized similarity by the feature of *followers*, *tags*, and *topics* between user i and j. $\eta_i (i = 1, 2, 3)$ is the weight value indicating the significance when measuring attribute similarity. Here, η_i will be determined by the significance degrees of the features mentioned in above subsection, then we empirically assign $\eta_1 : \eta_2 : \eta_3 = 22.6 : 3.8 : 3.6$. $dist(i, j)$ is the interest divergence between user i and j defined in Eq. 8.

By applying our new approach introduced above, the information of link structure $\{L_{ij}\}$ and content similarity $\{S_{ij}\}$ can be unitedly converted to the edge weight of a remodeled network $\{W_{ij}\}$, which is undirected and weighted. Then, many well-developed community detection methods for weighted networks can be applied to microgroup detection on Sina weibo, without losing considerations of link structure and user content.

In this paper, three well-known algorithms of CNM [6], Infomap [37], and OSLOM [28] are employed to identify communities on our remodeled network. CNM is a fast greedy modularity optimization algorithm proposed by Clauset, Newman, and Moore, and its key idea is starting from a set of isolated nodes, and the links of the original network are iteratively added to produce the largest possible increase in the modularity degree at each step. Infomap is a new information theoretic approach proposed by Rosvall and Bergstrom that reveals community structure in weighted networks, and the key is to decompose a network into modules by optimally compressing a description of information flows on the network. *Order Statistics Local Optimization Method* (OSLOM) is the first method capable of detecting communities in networks accounting for edge directions, edge weights, and overlapping communities. The method is based on the local optimization of a fitness function expressing the statistical significance of communities with respect to random fluctuations, which is estimated with the tools of Extreme and Order Statistics.

6 Comparing partitions

Many community detection methods have been presented based on a range of different ideas. In order to estimate the performance of these methods, the best way is to compare the outcomes of these methods with real community partition on artificial or real-word networks, and this can be done using *similarity measures* as mentioned in Refs. [15,41]. Newman used the *fraction of correctly identified nodes* to measure the goodness of community detection algorithm in Ref. [17]; however, it does not work well in some cases, then some other measurements have been proposed, in which the measurement of *Normalized Mutual Information* (*NMI*) borrowed from information theory has been proved to be reliable [8], and will be adopted to estimate the performance of our united method.

The measurement of *Normalized Mutual Information* is based on defining a confusion matrix $[N_{ij}]$, where the rows correspond to the "real" communities, and the columns correspond to the "detected" communities. The member of $[N_{ij}]$, N_{ij} is the number of nodes in the "real" community i that appear in the "detected" community j. The number of "real" communities is denoted c_A and the number of "detected" communities is denoted c_B, the sum over row i of matrix $[N_{ij}]$ is denoted N_i, and the sum over column j is denoted $N_{.j}$. N is the sum of all members of the matrix $[N_{ij}]$. Then, a measurement of similarity between the two partitions A and B, based on information theory, is defined as

$$NMI(A, B) = \frac{2I(A, B)}{H(A) + H(B)} \tag{16}$$

Table 2 Basic information of the experimental networks

Network	Nodes	Edges	Microgroups
TU	3016	18434	8
HUST	1084	8913	9
BNU	1572	17581	10
RUC	1617	10691	5

where

$$I(A,B) = -2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)$$

$$H(A) = \sum_{i=1}^{c_A} N_{i.}\log\left(\frac{N_{i.}}{N}\right), \quad H(B) = \sum_{j=1}^{c_B} N_{.j}\log\left(\frac{N_{.j}}{N}\right)$$

NMI takes the maximum value of 1 if the "detected" partition is completely consistent with the "real" case, whereas it has an expected value of 0 if the two partitions are totally independent. The measurement of NMI is currently often used for performance estimation of community detection algorithms.

7 Experiments and results

In this section, we validate the effectiveness and efficiency of our new proposed microgroup detection method by applying it to several real-world networks from Sina weibo. Since almost all of the frequently used test networks (e.g., Zachary [49], Football [17], Dolphins [30]) in community detection have no information about user attribute, we collect four collegiate social networks from Sina weibo as our experimental networks, and each of them is composed by some microgroups. We compare the detected communities using our method with the truth community structures of the sampled networks. Table 2 shows the basic information of our four experimental networks, and the details are described as follows.

- TU network, a social network of Sina weibo users from TSinghua University, one of the most famous universities in China. The network is composed by eight microgroups, namely Fine Arts Institute, Architecture Institute, EMBA Club, Chinese Institute, Law Institute, TSinghua Library, Electronic Engineering Institute, and Industrial Engineering Institute.
- HUST network, a social network of Sina weibo users from Huazhong University of Science and Technology, a famous university located in Wuhan of China and is composed by nine microgroups, namely Life Science Institute, Wenhua Institute, Management Institute, Wuchang Branch School, Alumnus in Guangzhou, Alumnus in Shanghai, Architecture Institute, Communication Institute, and Software Institute.
- BNU network, a social network of Sina weibo users from Beijing Normal University, a famous university located in Beijing of China and is composed by ten microgroups, namely Communication Institute, Digital Media Institute, Law and Politics Institute, Humanities Institute, Zhuhai Branch School, Sliding Wheel Club, Foreigner Student Club, Basketball Club, Youth Union, and HR Manager Club.

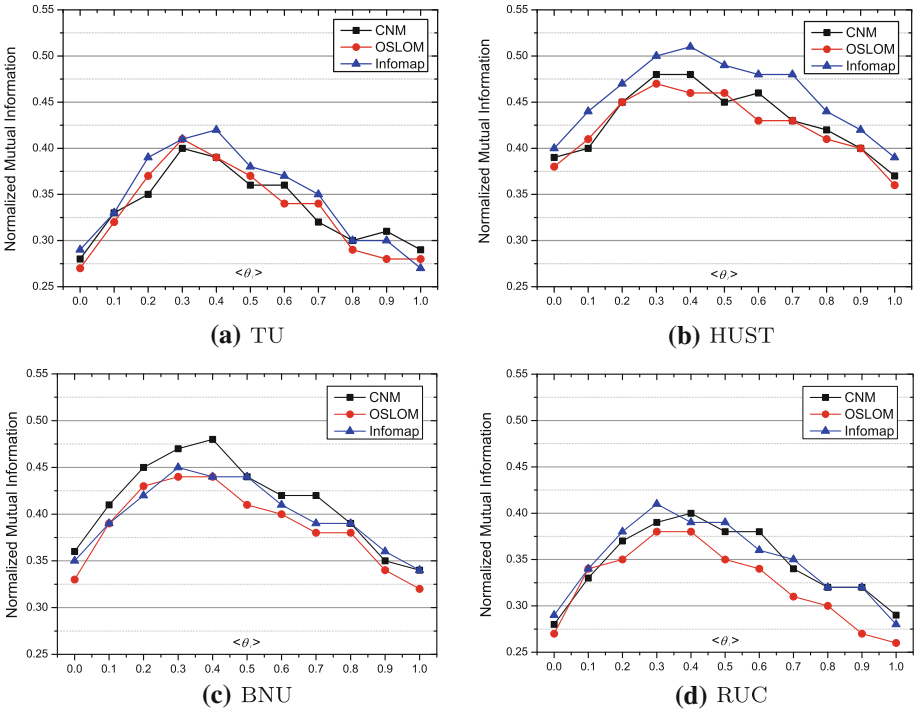

Fig. 8 Results of microgroup detection considering only user content on four real-world networks. **a** TU, **b** HUST, **c** BNU, and **d** RUC

– RUC network, a social network of Sina weibo users from Renmin University of China, a famous university located in Beijing of China and is composed by five microgroups, namely School of Journalism, MBA Club, Law Institute, On-job Postgraduates, and Economics Institute.

By applying our method, we convert the link relationship and content similarity between two users to the edge weight of a new generated network with Eq. 12. As the weight value of α and β, respectively, indicates the importance of the link and content in microgroup detection, we gradually vary the two values (α, β) to seek a better assignment in our experiments. Then, the community detection algorithms support for undirected and weighted network can be employed on the new remodeled network.

When measuring the content similarity between two users using Eq. 15, the two weight values (θ_1 and θ_2, $\theta_1 + \theta_2 = 1$) have not been assigned yet. That is, we have no idea on which factor is more important in measuring the content similarity. Thus, for seeking an appropriate assignment for the two weight values, we firstly ignore the link information, and only consider the user content for microgroup detection on test networks. Figure 8 illustrates the results of microgroup detection considering only content similarity on four real-world networks. The weight value θ_1, i.e., the significance degree of attribute similarity, varies from 0 to 1 with 0.1 as the step size, and with the increase of θ_1, the performance for three microgroup detection algorithms firstly becomes higher and then decreases in the latter case. When θ_1 is about 0.3 or 0.4, the performance nearly reaches the highest level on four test networks. For this reason, we conclude that individual interest extracted from published posts

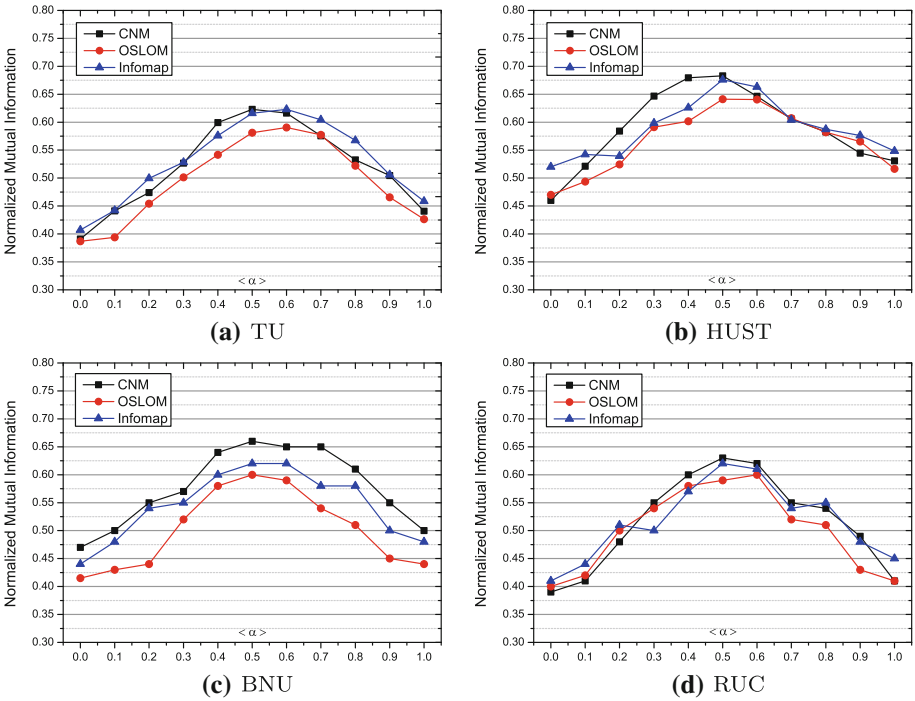

Fig. 9 Results of our united method on four real-world networks. **a** TU, **b** HUST, **c** BNU, **d** UC

is more significant than user attribute on measuring the content similarity between two users, and empirically set $\theta_1 = 0.3, \theta_2 = 0.7$ in our work.

In Fig. 9, we show the experimental results of our method on four test networks. In order to reveal the performance of our method under different parameters, the weight value of link structure (α in Eq. 12, $\beta = 1 - \alpha$) is tuned from 0 to 1 with 0.1 as the step size, and $\alpha = 0$ is the case of conventional clustering methods considering only user attribute, $\alpha = 1$ means the case of traditional community detection methods based only on link structure. On each test social network, the tendencies of tree curves, respectively, for CNM, OSLOM, and Infomap are very consistent. The performance of three methods is sensitive to α, increases with the growth of α, and then achieves the highest level when α is about 0.5, after that, the performance decreases with the growth of α. In addition, by comparing the three curves, we can know that the methods of CNM and Infomap always gain better outcomes than OSLOM. Thus, from the comparing results, we can conclude that, no matter what algorithms you employ, link structure and user content play almost equally important role for microgroup detection on Sina weibo, and our united method considering both aspects allows us to get better performance than traditional algorithms based on either link relationship or content similarity.

As an excellent approach for community detection on text citation networks, the discriminative model proposed by Yang et al. [48] was compared with our united model taking CNM algorithm as an example. When classifying individuals in microblogging networks using Yang's method, the link denotes the "following" relationship, and the content indicates the published posts. We may note that the implementing code of Yang's method can be downloaded from "http://www.cse.msu.edu/~yangtia1/codes/community_detection.zip".

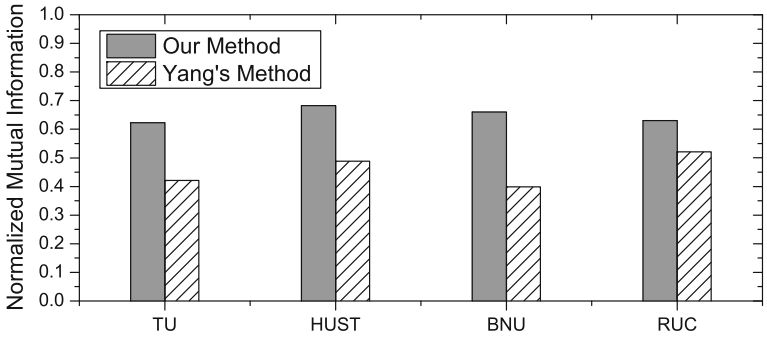

Fig. 10 Comparison of Yang's method and our united method on four real-world networks

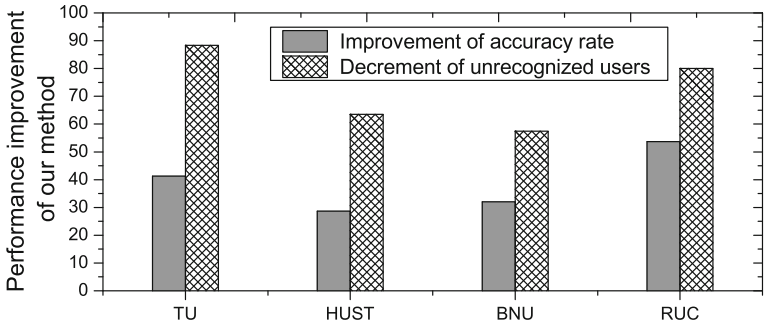

Fig. 11 Performance improvement of our method on four real-world networks

Figure 10 shows the comparing results of Yang's method and our united method for community detection on four real-world networks, from which we can see that our method outperforms Yang's method in all real-world networks. The superiority of our united method could be explained by the difference of the main attention of those two methods. In our work, we mainly focus on the problem of community detection in microblogging system and consider interest similarity as a significant factor to classify individuals; then, many elements are exploited to measure individual interest by analyzing the content including published posts, labeled tags, and discussed topics. However, Yang et al. mainly concern the problem of community detection on text citation networks, which is very similar to the problem of text classification, then they only consider the "citing" relationship and content similarity between text nodes, but ignore human features. Hence, we say that Yang's method may work well for community detection on text citation network, but not for that on microblogging network.

Finally, Fig. 11 shows the accuracy improvement of our united method compared with the method based only on link structure, taking CNM algorithm as an example. On four test social networks, the average accuracy improvement indicated by left pillars in Fig. 11 is about 39 %, and the biggest is 53 % on RUC network. Furthermore, most microgroups on Sina weibo are very sparse and many isolated users cannot be clustered to any community using the methods based only on link structure. Fortunately, using our method, more users can be identified. The right pillars in Fig. 11 show the decrement rates of unrecognized users on test networks, and the mean value is about 72 %.

8 Conclusion and future work

In this paper, we proposed a united method to combine link structure and user content for microgroup detection on Sina weibo, the most popular microblogging system in China. By analyzing the topological characteristics and content similarity of the sampled microgroups, we have observed: unlike many traditional social networks, degree assortativities on most microgroups are negative or weakly positive, which implies that most users with few followers tend to follow those with many followers, and the "following" relationship on Sina weibo does more in sharing information than maintaining friendship. From the negative assortativity by many user attributes, we believe that the two ends with link relationship on Sina weibo are not so similar with each other like those on other OSNs (e.g., Facebook, Flikr). Also, from another point of view, we computed the average similarity of members from the same microgroup based on different attributes including *followers list*, *followings list*, *tags list*, and *topics list*. By comparing with the random case, we found that the feature of *followers* is the most significant for microgroup detection, followed by the features of *tags* and *topics*, while the feature of *followings* is nearly indistinctive with random case. Furthermore, we extracted individual interest from published posts using LDA model and measured the average interest similarity between users from the same microgroup. By comparing with random case, we observed that users from the same microgroup tend to share more common interests.

Based on our observations on the characteristics of the sampled microgroups, we proposed a united method to remodel the network for microgroup detection on Sina weibo. Using our method, link structure and content similarity between two users are converted to the edge weight of a new remodeled network. Through extensive experiments on four real-world social networks with known community structures, we observed that our method obtains significant improvement over the traditional community detection algorithms considering either link structure or user content.

As more and more microblogging systems emerge on the Internet, our work of microgroup detection on Sina weibo can also be applied to solve the similar problems on many other microblogging networks like Twitter and Google+. For future work, we will validate the performance of our method on more data sets.

Acknowledgments We thank anonymous reviewers for their useful comments and suggestions. This work was partially supported by the fund of open project from the State Key Lab of Software Development Environment, China (No. SKLSDE-2011KF-06), the National High Technology Research and Development Program of China (863 Program) (No. 2012AA011005), and the State Key Laboratory of Mathematical Engineering and Advanced Computing, China. Part of this research was done when the first author visited the State Key Lab of Software Development Environment, Beihang University, China. We would like to thank Dr. Jichang Zhao, Dr. Xu Feng, and Dr. Xiao Liang for their encouragement and support.

References

1. Agrawal R, Gehrke J, Gunopulos D, Raghavan P (1998) Automatic subspace clustering of high dimensional data for data mining applications. In: SIGMOD conference'98. pp 94–105
2. Andreopoulos B, An A, Wang X, Schroeder M (2009) A roadmap of clustering algorithms: finding a match for a biomedical application. Brief Bioinform 10(3):297–314
3. Arenas A, Díaz-Guilera A, Pérez-Vicente CJ (2006) Synchronization reveals topological scales in complex networks. Phys Rev Lett 96(11):114102
4. Cha M, Mislove A, Gummadi PK (2009) A measurement-driven analysis of information propagation in the flickr social network. In: World wide web conference series, pp 721–730
5. Cheeseman P, Stutz J (1996) Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, Menlo Park, CA

6. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066111
7. Cohn DA, Hofmann T (2001) The missing link—a probabilistic model of document content and hypertext connectivity. In: Leen TK, Dietterich TG, Tresp V (eds) Advances in Neural information processing systems 13. MIT Press, pp 430–436
8. Danon L, Duch J, Arenas A, Daz-guilera A (2005) Comparing community structure identification. J Stat Mech Theory Exp 9008:09008
9. Dietz L, Bickel S, Scheffer T (2007) Unsupervised prediction of citation influences. In: Proceedings of the 24th international conference on machine learning, pp 233–240
10. Duan D, Li Y, Jin Y, Lu Z (2009) Community mining on dynamic weighted directed graphs. In: Proceedings of international conference on information and knowledge management, pp 11–18
11. Endres DM, Schindelin JE (2003) A new metric for probability distributions. IEEE Trans Inf Theory 49(7):1858–1860
12. Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD'96, pp 226–231
13. Flake G, Lawrence S, Giles C, Coetzee F (2002) Self-organization and identification of Web communities. Computer 35(3):66–70
14. Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174
15. Fortunato S, Castellano C (2007) Community structure in graphs. eprint arXiv: 0712.2716
16. Getoor L, Friedman N, Koller D, Taskar B (2002) Learning probabilistic models of link structure. J Mach Learn Res 3:679–707
17. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. PNAS 99(12):7821–7826
18. Gregory S (2010) Finding overlapping communities in networks by label propagation. New J Phys 12(10):103018+
19. Gruber A, Rosen-Zvi M, Weiss Y (2008) Latent topic models for hypertext. In: McAllester DA, Myllymäki P (eds) Proceedings of the 24th conference in uncertainty in artificial intelligence (UAI-08). AUI Press, Corvallis, Oregon, pp 230–239
20. Hochbaum DS, Shmoys DB (1985) A best possible heuristic for the k-center problem. Math Oper Res 10(2):180–184
21. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31:(3)264–323
22. Kalogeratos A, Likas A (2011) Document clustering using synthetic cluster prototypes. Data Knowl Eng 70(3):284–306
23. Kernighan BW, Lin S (1970) An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J 49(1):291–307
24. Kim Y, Son SW, Jeong H (2009) Community identification in directed networks. In: Zhou J (ed) Complex sciences, vol 5 of lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, pp 2050–2053
25. Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. In: Eliassi-Rad T, Ungar LH, Craven M, Gunopulos Dimitrios (eds) Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, pp 611–617
26. Kwak H, Lee C, Park H, Moon SB (2010) What is Twitter, a social network or a news media? In: World wide web conference series, pp 591–600
27. Lai D, Lu H, Nardini C (2010) Finding communities in directed networks by pagerank random walk induced network embedding. Physica A Stat Mech Appl 389:2443–2454
28. Lancichinetti A, Radicchi F, Ramasco JJ (2010) Statistical significance of communities in networks. Phys Rev E 81(4):046110
29. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58:1019–1031
30. Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E (2003) The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behav Ecol Sociobiol 54(4):396–405
31. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67(2):026126
32. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
33. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814
34. Pothen A, Simon HD, Liou K-P (1990) Partitioning sparse matrices with eigenvectors of graphs. SIAM J Matrix Anal Appl 11(3):430–452

35. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci 101(9):2658
36. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106
37. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. PNAS 105:1118
38. Stanoev A, Smilkov D, Kocarev L (2011) Identifying communities by influence dynamics in social networks. CoRR abs/1104.5247. http://arxiv.org/abs/1104.5247
39. Stephen EE, Fienberg S, Lafferty J (2004) Mixed membership models of scientific publications. Proc Natl Acad Sci 101(suppl 1):5220–5227. doi:10.1073/pnas.0307760101
40. Topsoe F (2000) Some inequalities for information divergence and related measures of discrimination. IEEE Trans Inf Theory 46(4):1602–1609
41. Traud AL, Kelsic ED, Mucha PJ, Porter MA (2009) Comparing community structure to characteristics in online collegiate social networks. In: Proceedings of the 2009 APS March meeting
42. Wang X, Tang L, Liu H, Wang L (2012) Learning with multi-resolution overlapping communities. Knowl Inf Syst 1–19. doi:10.1007/s10115-012-0555-0
43. White S, Smyth P (2005) A spectral clustering approach to finding communities in graphs. Proc SIAM Int Conf Data Min
44. Xiang R, Neville J, Rogati M (2010) Modeling relationship strength in online social networks. In: Rappa M, Jones P, Freire J, Chakrabarti S (eds) WWW. ACM, pp 981–990
45. Xiong X, Niu X, Zhou G, Xu K, Huang Y (2011) Microgroup mining on tsina via network structure and user attribute. In: Tang J, King I, Chen L, Wang J (eds) ADMA (2), vol 7121 of lecture notes in computer science. Springer, pp 138–151
46. Yan F, Cai S, Zhang M, Liu G, Deng Z (2013) A clique-superposition model for social networks. Sci China Inf Sci 56(5):52113. doi:10.1007/s11432-011-4526-y
47. Yang T, Chi Y, Zhu S, Gong Y, Jin R (2010) Directed network community detection: a popularity and productivity link model. In: SIAM international conference on data mining, pp 742–753
48. Yang T, Jin R, Chi Y, Zhu S (2009) Combining link and content for community detection: a discriminative approach. In: Knowledge discovery and data mining. pp 927–936
49. Zachary W (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33:452–473
50. Zhang K, Lo D, Lim E-P, Prasetyo P (2012) Mining indirect antagonistic communities from social interactions. Knowl Inf Syst 1–31. doi:10.1007/s10115-012-0519-4
51. Zhang T, Ramakrishnan R, Livny M (1997) Birch: a new data clustering algorithm and its applications. Data Min Knowl Discov 1(2):141–182
52. Zhao J, Wu J, Feng X, Xiong H, Xu K (2012) Information propagation in online social networks: a tie-strength perspective. Knowl Inf Syst 32(3):589–608

Author Biographies

Xiaobing Xiong is currently a Ph.D. candidate at the State Key Laboratory of Mathematical Engineering and Advanced Computing, China. He received a B.E. degree and M.E. degree from the National Digital Switching System Engineering and Technological Research Center in 2006 and 2009, respectively. From 2010 to 2011, he did research work at the State Key Lab of Software Development Environment, Beihang University, China. His research interests include data mining and social network analysis.

Gang Zhou is currently an Associate Professor at the State Key Laboratory of Mathematical Engineering and Advanced Computing, China. He received a M.E. degree and Ph.D. degree from Beihang University in 1999 and 2007, respectively. His research interests include data mining, database, social network analysis, and distributed system.

Xiang Niu received a B.E. degree from University of Science and Technology Beijing, Beijing, China, in 2008. He is currently a postgraduate student at the Department of Computer Science, Beihang University, China. His research interests include data mining, multimedia, and online social network.

Yongzhong Huang is currently a Professor at the State Key Laboratory of Mathematical Engineering and Advanced Computing, China. He received his M.E. degree and Ph.D. degree from the National Digital Switching System Engineering and Technological Research Center. His research interests include data mining, distributed system, and database.

Ke Xu is currently a Professor at Beihang University, China. He received his B.E., M.E., and Ph.D. degrees from Beihang University in 1993, 1996, and 2000, respectively. His research interests include algorithm and complexity, data mining, and complex networks.