Neurocomputing 179 (2016) 219–227

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An improved density peaks-based clustering method for social circle discovery in social networks

Mengmeng Wang [a,b], Wanli Zuo [a,b], Ying Wang [a,b,c,*]

[a] College of Computer Science and Technology, Jilin University, Changchun 130012, PR China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun 130012, PR China
[c] College of Mathematics, Jilin University, Changchun 130012, PR China

ARTICLE INFO

Article history:
Received 18 May 2015
Received in revised form
6 November 2015
Accepted 29 November 2015
Communicated by Yongdong Zhang
Available online 15 December 2015

Keywords:
Discovering overlapping social circles
Improved density peaks-based clustering method
In-link Salton metric
Out-link Salton metric
Social networks

ABSTRACT

With the development of Internet, social networks have become important platforms which allow users to follow streams of posts generated by their friends and acquaintances. Through mining a collection of nodes with similarities, community detection can make us understand the characteristics of complex network deeply. Therefore, community detection has attracted increasing attention in recent years. Since the problem of discovering social circles is posed as a community detecting problem, hence, in this paper, targeted at on-line social networks, we investigate how to exploit user's profile and topological structure information in social circle discovery. Firstly, according to directionality of linkages, we put forward in-link Salton metric and out-link Salton metric to measure user's topological structure. Then we propose an improved density peaks-based clustering method and deploy it to discover social circles with overlap on account of user's profile- and topological structure-based features. Experiments on real-world dataset demonstrate the effectiveness of the proposed framework. Further experiments are conducted to understand the importance of different parameters and different features in social circle discovery.

1. Introduction

With the advent of online social networks, social network analysis has gradually become a hot issue in both academia and industry. Since community structure, which often represents specific organized groups of users with similar attributes, hobbies or closer relationships [1], is a significant property of social networks, so community detection is one of the basic research problems in social network analysis. Detecting community is very important for understanding the characteristics of complex network, discovering latent topology, predicting network evolution and so on [2]. Besides, identifying community structure can facilitate many tasks such as following/follower recommendation [3], task allocation [4], proximity alignment [5], maximizing influence [6], retweeting behavior prediction [7], mining cybercriminal networks [8] and so forth. A novel function has been provided in some major social networks: users can categorize their friends into social circles which can be used to filter status updates posted by distant acquaintances, hide personal information from coworkers and share groups of users that others may wish to follow [9]. Therefore, social circle discovery which can fall into the domain of community detection has attracted increasing attention in recent years. As stated, our work on social circle discovery is motivated by its broad application prospect.

The purpose of community detection is to find a group of users with similar ideas, beliefs, motivations or other common features so as to better understand social networks. However, most of existing community detection approaches which only considered structural features (e.g., links) [10] may ignore much information that associated with community, such as user's background information and interaction information [11]. Besides, isolating user's relationship with user's contents may result in finding unreasonable community structure, while most community discovery algorithms which considered both types of information were usually complex. Hence, in this paper, we propose an improved density peaks-based clustering method which incorporates both structural and attribute information of users for social circle discovery in social networks (denoted as DPSCD), and our main contributions are summarized next.

(1) Put forward in-link Salton metric and out-link Salton metric according to directionality of linkages to achieve a better representation on adjacent degree between users in directed social networks;

(2) Improve a fast clustering method with novel density estimator and extra social circle integration step in order to better

* Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun 130012, PR China. Tel.: +86 13604413931.
E-mail addresses: wmmwwlh@126.com (M. Wang), zuowl@jlu.edu.cn (W. Zuo), wangying2010@jlu.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.11.091
0925-2312/

adapt to large statistical errors, followed by employing it to detect overlapping social circles in social networks; and

(3) Evaluate DPSCD on real-world dataset Facebook, Google+, Twitter and elaborate the importance of different parameters and different features on social circle discovery.

The rest of paper is organized as follows: Section 2 describes the related work; Section 3 defines the method we propose; Details of the experimental results and dataset which is used in this study are given in Section 4. Finally conclusion appears in Section 5.

2. Related work

It has been pointed out that social circle discovery can be formulated as a community detection problem on a user's ego network (the network of friendships between his/her friends) [9–12]. A great deal of works have been done on community detection, Newman and Girvan [13] first defined a community as a subgraph containing nodes which were more densely linked to each other within such subgraphs and sparse between them. Community discovery provides an important means to understand the structure of complex networks deeply [14], therefore, the issue of detecting community in social networks has received increasing attention in recent years. Existing community detection algorithms can be roughly categorized into three groups: relationship-based method, content-based method and comprehensive method where users' relationships and contents are merged with users' attributes.

Since topological structure can influence individuals' behaviors in the network [15], as a consequent, discovering communities based on users' relationships is still a mainstream method. In relationship-based algorithms, through utilizing properties of relationships, communities with closer internal relationships are detected. Li et al. [16] first proposed two new algorithms based on evolutionary algorithm and clonal selection algorithm (denoted as EA-SN and CSA-SN, respectively). Then, they integrated a hill climbing (HC) strategy into EA-SN and CSA-SN to form two new memetic algorithms (EAHC-SN and CSAHC-SN). The experimental results not only showed the capability and high efficiency of EAHC-SN and CSAHC-SN in successfully detecting communities from signed networks, but also indicated that both the two objective functions (improved modularity and improved modularity density) were efficient to some extent. Ma et al. [17] presented a seed insensitive method for local community detection which estimated similarity among vertices by investigating vertices' neighborhoods and revealed a local community by maximizing its internal similarity and minimizing its external similarity simultaneously. Extensive experimental results on both synthetic and real-world data sets verified the effectiveness of the proposed algorithm. Cai et al. [18] proposed a novel discrete PSO algorithm for identifying community structures in signed networks. In order to make PSO be proper for discrete scenarios, they redesigned particles' status in a discrete form, followed by reformulating particles' updating rules through making use of topological structure of signed network. Extensive experiments demonstrated that the proposed method was effective and promising. Li et al. [19] first improved nonnegative matrix factorization method with modeling network as a weighted directed graph and using diagonally dominant matrix as constraint condition to obtain community membership of each node as well as interactions between communities. The results demonstrated that the proposed method was useful and applicable both in weighted directed model and undirected model for community discovery over other related matrix factorization methods. Rhouma et al. [20] proposed an overlapping community detecting algorithm called DOCNet (Detecting overlapping communities in Networks) which was based on local optimization of a fitness function and a fuzzy belonging degree of different nodes. The main strategy of this algorithm was to find an initial core and add suitable nodes to expand it until a stopping criterion was met. Experimental results demonstrated that DOCNet was efficient and highly reliable for detecting overlapping groups, compared with four newly known proposals. However, DOCNet model cannot adapt to weighted and directed networks. Qiu et al. [21] first generated a probability transition matrix by applying random walk to a social network, followed by training a Gaussian mixture model using the matrix. And then, overlapping communities were derived by analyzing mean vectors of the Gaussian mixture model. The experiments conducted on synthetic and real dataset demonstrated the feasibility and applicability of the proposed algorithm. Instead of using eigenvectors in spectral clustering algorithms, Huang et al. [22] put forward a regularized spectral clustering algorithm which chose sample matrix of social network to construct a target function that can partition social network naturally. The experiments shown the proposed method achieved good results with relatively smaller computational cost compared to spectral clustering algorithm. Wu et al. [23] introduced a cosine-pattern-based community extraction framework. It first extracted the so-called asymptotically equivalent structures (AESs) from networks, from which nodes were further partitioned into crisp communities using any of existing methods. A novel cosine-pattern mining algorithm based on the ordered anti monotone of cosine similarity was thus proposed for the efficient extraction of AESs. Experiments on various real-world social networks demonstrated the advantage of extracting view of community detection.

However, Dang and Viennet [24] pointed out that in real-world networks, in addition to topological structure (i.e., links), content information was also available. Besides, considering network structural information only may fail to detect interpretable overlapping communities since structural information of online social networks is often sparse and weak. Sang and Xu [25] held the view that the social links were well recognized forces that govern the behaviors of involved users as well as the dynamics of social networks. Additionally, through splitting all the Flickr user pairs into two parts, i.e., with relations and without relations, respectively and calculating the average of the common contact number, common interested group number and tag-based similarity in the respective user pairs, Yan et al. [26] pointed out that users generally had more common contacts and common interested groups with their friends than other people. Besides, friends may also use more similar tags in their uploaded images which may be influenced by each other. So some scholars began to calculate distance/similarity between users on their generated information.

Yin et al. [27] incorporated community discovery into topic analysis in text-associated graphs and proposed a community-based topic analysis framework called LCTA (Latent Community Topic Analysis). The proposed framework handled both topic modeling and community discovery to guarantee the topical coherence in the communities so that users in the same community were closely linked to each other and shared common latent topics. They compared different methods and performed extensive experiments on two real datasets. The results confirmed the hypothesis that topics could help understand community structure, while community structure could help model topics. Taking Flickr as one exemplary social media platform, Zhuang et al. [28] found that the taste/interest of a user in photos can be implicitly mined from the photos uploaded by the user and they proposed a content-aware low-rank matrix recovery technique for community discovery. First, they modeled the observed indicator matrix of the Flickr community as a summation of a low-rank true matrix and a sparse error matrix. And then, they formulated an optimization

problem by regularizing the true matrix to coincide with the available rich context and content (i.e., photos and their associated tags). Finally, they conducted extensive experiments on a large-scale real dataset from Flickr and presented insights about Flickr community. However, Zhuang et al. mainly exploited the textual and visual information, but in Flickr, there were still much other heterogeneous metadata available other than the images and tags.

Nevertheless, it cannot make a comprehensive analysis on overlapping communities based on specific types of information only, consequently, some efforts have been made to integrate users' structural, content and attribute information. The main idea of comprehensive algorithm is to design a distance/similarity measure for vertex pairs that combines structural, content and attribute information of nodes. Based on this measure, standard clustering algorithms such as k-medoids and spectral clustering are then applied to cluster nodes [29]. To avoid presetting the number and the size of communities, Xin et al. [30] presented a clustering algorithm for community detection based on a link-field-topic (LFT) model which separated community-topic detection into LDA sampling and semantic community detection. Through experimental analysis, LFT model approached to the optimum of all classical semantic community detection algorithms. Hu and Yang [12] proposed an enhanced link clustering method to identify social circles on ego networks. Through constructing an edge profile for each edge, they integrated node profile and network structure. Experiments on several real datasets demonstrated that the proposed method was not only effective, but also more efficient in comparison with state-of-the-art methods when taking edge similarity instead of node similarity to discriminate nodes into different circles. Deitrick and Hu [31] used sentiment classification to enhance community detection. Firstly, community detection was performed on friend/follower networks of the four Microsoft accounts using SLPA and Infomap algorithms. Then, three types of additional features: replies, mentions, and retweets; hashtags; and sentiment classifications were used to iteratively increase edge weights in the four social networks, and community detection was repeated on the networks using edge weights updated with each day's data. The results revealed that modularity values were increased for the community partitions detected in three of the four networks studied by combining community detection and sentiment analysis. Dang and Viennet [24] studied the relationship between semantic similarity of users and topology of social networks (homophily concept). They proposed two structure-attribute clustering approaches (improved Louvain algorithm and k-NN method) to extract communities on several real-world datasets. Experimental results demonstrated that their methods provided more meaningful communities than conventional methods that considered relationship information only. McAuley [9] first predicted hard assignment of a node to multiple circles, which proved critical for good performance, followed by proposing a parameterized definition of profile similarity to learn the dimensions of similarity along which links emerged. And then, an unsupervised algorithm was devised to jointly optimize the latent variables and the profile similarity parameters so as to best explain the observed network data.

To sum up, community detection in social networks is in the stage of development, how to depict directionality of linkages, how to fuse multidimensional features reasonably and how to build model that could detect community efficiently can be very challenging jobs. To this end, we present an improved density peaks-based clustering method which is appropriate for overlapping social circle discovery in social networks.

3. Improved density peaks-based clustering method for overlapping social circle discovery

3.1. Problem formulation

In this paper, we formally define social circle discovery as: given a set of ego-networks $G=\{G_1, G_2, \cdots, G_k, \cdots, G_n\}$where n encodes the number of ego-networks, $G_k=(V_k, E_k)$ encodes user k's ego-network (a network of relationships between k's friends) where V_k and E_k encode the set of users and the set of edges in G_k (user k is not included since creators of circles do not themselves appear in their own circles [9]), additionally, each user is associated with an attribute vector, we aim to predict a set of circles (a circle represents a set of users) $C(k)=\{C_1(k), C_2(k), \cdots, C_j(k)\}$ (j encodes the number of circles that user k may create, $C_j(k)$ encodes users in the j-th circle that user k may create) for each ego-network G_k via profile and topological structure information.

3.2. Social circle discovery features definition

Mislove et al. [32] found that nodes in one circle represented transitionally similarity, but not necessary to be very similar to each other or densely connected and each circle's members usually shared common properties or traits. Consequently, merely based on topological structure cannot make a comprehensive analysis on social circles. Thus, in this paper, we synthesize profile and topological structure information so as to achieve higher accuracy in social circle discovery.

3.2.1. Profile-based features

Silva et al. [33] pointed out that rich information was encoded in the content of networks such as node content. Thus, in this paper, according to ground-truth data from three major social networking sites: Facebook, Google+, and Twitter[1] which will be introduced in detail in Section 4.1, we name 26 categories, including hometowns, birthdays, colleagues, political affiliations and so forth as profile-based features in Facebook dataset, gender, last name, job titles, institutions, universities, places lived as profile-based features in Google+ dataset, the set of hashtags and mentions used by each user during two-weeks' worth of tweets as profile-based features in Twitter dataset.

3.2.2. Topological structure-based feature

The linkage information between vertices plays a critical role to evaluate vertices' similarities [17]. Common Neighbors metric assumed that similarity between users was proportional to the number of their common neighbors [34]. On the basis of Common Neighbors metric, Salton metric introduced users' degree. Since both Google+and Twitter are directed social networks, in addition, non-reciprocal friendships, which may reflect moderately valued friendship ties [35], are more important than reciprocal friends, hence, we put forward in-link Salton metric and out-link Salton metric to measure user's network topological structure in directed social networks:

$$inSa(i,v)=\frac{|\Gamma^{in}(i)\cap\Gamma^{in}(v)|}{\sqrt{|\Gamma^{in}(i)|\times|\Gamma^{in}(v)|}} \tag{1}$$

$$outSa(i,v)=\frac{|\Gamma^{out}(i)\cap\Gamma^{out}(v)|}{\sqrt{|\Gamma^{out}(i)|\times|\Gamma^{out}(v)|}} \tag{2}$$

where $inSa(i,v)$ and $outSa(i,v)$ stand for in-link Salton metric and out-link Salton metric between user i and user v respectively; Γ^{in}

[1] http://snap.stanford.edu/data/

(i) and $\Gamma^{in}(v)$ stand for in-link users set of user i and user v respectively; $\Gamma^{out}(i)$ and $\Gamma^{out}(v)$ stand for out-link users set of user i and user v respectively, where in-link and out-link are defined by follower relationship; $|\bullet|$ stands for the number of elements in a set. For undirected social networks, we employ traditional Salton metric to measure user's network topological structure:

$$Sa(i,v)=\frac{|\Gamma(i)\cap\Gamma(v)|}{\sqrt{|\Gamma(i)|\times|\Gamma(v)|}} \tag{3}$$

where $Sa(i,v)$ presents Salton metric between user i and user v; $\Gamma(i)$ and $\Gamma(v)$ present neighbors set of user i and user v respectively, where neighbors are defined by undirected edges between users.

3.3. Algorithm for discovering overlapping social circles

Clustering algorithms are often used in community detection. Generally, an objective function is optimized iteratively in some clustering algorithms which may lead to lower efficiency. Rodriguez and Laio [36] proposed a fast clustering method that performed in a single step, which can recognize clusters regardless of their shape and of the dimensionality of the space where they were embedded. For its novel and concise idea, it provided a new train of thought for the design of clustering algorithm. Hence, we intend to investigate whether Rodriguez and Laio's method can be an effective method in social circle discovery. However, social circles are nested: small social circles build larger ones, which in turn group together to form even larger ones [37], the method Rodriguez and Laio put forward may fail to detect overlapping social circles and the density estimator may be unavoidably affected by large statistical errors for a small data set. Accordingly, in this paper, we do some improvements on Rodriguez and Laio's method so as to apply it in overlapping social circle discovery for the first time: we adopt Gaussian kernel as density estimator to avoid large statistical errors and detect overlapping social circles according to the distances between users in different social circles.

Our proposed method for discovering overlapping social circles consists of two stages: (1) stage of initial social circle clustering; (2) stage of social circle integration. The detailed descriptions are shown as follows.

3.3.1. Initial social circle clustering

In social networks, a few users master the whole net as core, while the others have little influence on the net, so different roles are played by users in community [19]. Hence, we first measure local density for each user in an ego-network via Gaussian kernel. User i's local density is calculated as follows:

$$\rho_i=\sum\nolimits_v exp\left(-\frac{\|d_{i,v}-d_c\|^2}{2\sigma^2}\right) \tag{4}$$

where σ represents a smoothing parameter, $\|d_{i,v}-d_c\|$ represents the Euclidean distance metric between $d_{i,v}$ and d_c, d_c represents a cutoff distance, $d_{i,v}$ represents the distance between user i and user v which is defined as:

$$d_{i,v}=\frac{1}{\alpha\times simP(i,v)+(1-\alpha)\times simN(i,v)} \tag{5}$$

where parameter α is introduced to control profile and topological structure information's contribution in user's distance, $simP(i,v)$ and $simN(i,v)$ denote profile similarity and topology structure similarity between user i and user v respectively. Similarity between users is inversely proportional to the distance between them. The calculation of $simP(i,v)$ is shown as:

$$simP(i,v)=\sum\nolimits_m P_i(m)\Delta P_v(m) \tag{6}$$

where $P_i(m)$ and $P_v(m)$ present the m-th profile-based feature's value of user i and user v respectively, Δ presents an operator, if $P_i(m)=P_v(m)$, then $P_i(m)\Delta P_v(m)=1$, and $P_i(m)\Delta P_v(m)=0$ otherwise. And the calculation of $simN(i,v)$ in directed social networks is shown as:

$$simN(i,v)=inSa(i,v)+outSa(i,v) \tag{7}$$

For undirected social networks, $simN(i,v)=Sa(i,v)$.

Secondly, we calculate distance from points of higher density for each user. User i's distance from points of higher density is calculated as:

$$\delta_i=\begin{cases} max_v(d_{i,v}) & \text{if user } i \text{ of the largest density} \\ min_{v:\rho_v>\rho_i}(d_{i,v}) & \text{otherwise} \end{cases} \tag{8}$$

And then, on the basis of the assumptions that cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density [36], from the decision graph in which x axis denotes values of local density and y axis denotes values of distance from points of higher density, only the users of high distance from points of higher density and relatively high local density are treated as cluster centers, while users who have a relatively high distance from points of higher density and a low local density can be considered as outliers.

Finally, we assign each remaining user to the same cluster as its nearest neighbor of higher density to achieve initial social circle clustering.

3.3.2. Social circle integration

A user may belong to multiple circles simultaneously [38], as a consequence, in order to detect overlapping social circles, consolidation has to happen within the initial clustered social circles. Furthermore, McAuley and Leskovec [9] found that 'stronger' circles will form within 'weaker' ones, e.g. a circle of friends from the same degree program may form within a circle from the same university, thus, in this phase, each user in the social circles where cluster centers of smaller values of local density will be modeled memberships to the social circles where cluster centers of larger values of local density. The social circle integration can be defined as Algorithm 1.

Algorithm 1 social circle integration

Input: user k's set of initial social circles $C'(k)=\{C_1'(k), C_2'(k),\cdots,C_j'(k)\}$ $(C_1'(k),C_2'(k),\cdots,C_j'(k)$ are sorted in ascending order according to value of local density of each social circle's cluster center)

Output: user k's set of final social circles

(1) For l from 1 to $j-1$ Do
(2) For each user $i\in C_l'(k)$ Do
(3) For each user $i'\in C_{l+1}'(k)$ Do
(4) If $d_{i,CC(l+1)}\le d_{i',CC(l+1)}$ Then
%$CC(l+1)$ denotes cluster center of $(l+1)$-th circle that user k may create
(5) $C_{l+1}'(k)\leftarrow i$
(6) Break
(7) End if
(8) End for
(9) End for
(10) End for
(11) Return $C'(k)$

3.4. Time complexity

Assume that the average number of users in an ego-network is N_e, the account of social circles is N_c and the average number of users in a social circle is N_{cu}. The complexity of DPSCD is analyzed as follows. In the stage of initial social circle clustering, calculating distance between users will take $O\left(\frac{N_e(N_e-1)}{2}\right)$ time; calculating

local density for users in the ego-network takes $O(N_e(N_e-1))$ time; it is taking $O\left(\frac{N_e(N_e-1)}{2}\log_2\frac{N_e(N_e-1)}{2}\right)$ time for sorting the set of distance between users and $O(N_e\log_2 N_e)$ time for sorting the value set of local density; calculating distance from points of higher density for users in the ego-network will take $O(N_e)$ time in the best case and $O\left(\frac{N_e^2(N_e-1)}{2}\right)$ time in the worst case, then finding initial social circle for each user is taking $O(N_e)$ time. So the best overall complexity of the first stage is $O\left(\frac{3N_e(N_e-1)}{2}\right)$ and the worst overall complexity of the first stage is $O\left(\frac{N_e^2(N_e-1)}{2}\right)$. In the stage of social circle integration, the best computation time of membership modeling process takes $O\left(\frac{N_{cu}N_c(N_c-1)}{2}\right)$ time and the worst computation time of membership modeling process takes $O\left(\frac{N_{cu}^2N_c(N_c-1)}{2}\right)$ time. Assume that the total number of ego-networks is N_t. Since $N_e \gg N_c$ and $N_e > N_{cu}$, hence, the best overall complexity of our proposed method is $O\left(N_t\left(\frac{3N_e(N_e-1)}{2}\right)\right)$ and the worst overall complexity of our proposed method is $O\left(N_t\left(\frac{N_e^2(N_e-1)}{2}\right)\right)$.

4. Experimental evaluation

In this section, we conduct experiments to assess the effectiveness of the proposed framework DPSCD. Through the experiments, we aim to answer the following two questions:

(1) How effective is the proposed framework, DPSCD, compared with other methods of social circle discovery?
(2) What are the effects of different parameters and different features on the performance of DPSCD?

4.1. Dataset

To study the problem of social circle discovery, we leverage a dataset of 1143 ego-networks and 5636 hand-labeled ground-truth circles from Facebook (undirected), Google+ (directed), and Twitter (directed) [9] to evaluate validity of the proposed method. Facebook data is collected through conducting a survey of ten users, who are asked to manually identify all the circles to which their friends belong. Examples of such circles include students of common universities, sports teams, relatives, and so forth. Facebook data is fully labeled, in the sense that every circle that a user considers to be a cohesive community is obtained. Google+ and Twitter data are collected from users who have shared at least two circles, and whose network information are publicly accessible, so Google+ and Twitter data are only partially labeled. Statistics of the dataset are shown in Table 1.

4.2. Evaluation metrics

In this paper, we utilize Balanced Error Rate (BER) and F1-score which are used in [9] and [12] as metrics so as to compare our work to the works in [9] and [12]. In details, the BER between predicted circle C and ground-truth circle $\overline{C}$ can be formulated as:

$$BER\left(C,\overline{C}\right)=\frac{1}{2}\left(\frac{|C/\overline{C}|}{|C|}+\frac{|\overline{C}/C|}{|\overline{C}|}\right) \tag{9}$$

where / denotes a difference-set operator. Since false positives and false negatives are equally important in BER, thus, trivial or random predictions will incur extremely low errors (an error of 0.5 average). And the F1-score is calculated as:

$$F_1\left(C,\overline{C}\right)=\frac{2\times p\left(C,\overline{C}\right)\times r\left(C,\overline{C}\right)}{p\left(C,\overline{C}\right)+r\left(C,\overline{C}\right)} \tag{10}$$

where $p\left(C,\overline{C}\right)$ represents precision of C to $\overline{C}$ which is defined as:

$$p\left(C,\overline{C}\right)=\frac{|C\cap\overline{C}|}{|C|} \tag{11}$$

and $r\left(C,\overline{C}\right)$ represents recall of C to $\overline{C}$ which is defined as:

$$r\left(C,\overline{C}\right)=\frac{|\overline{C}\cap C|}{|\overline{C}|} \tag{12}$$

In this paper, we evaluate the performance of the proposed framework via the optimal matches of BER and F1-score for that we are short on the knowledge about the correspondence between C and $\overline{C}$. The optimal matches of BER and F1-score are defined as:

$$\max_{f:C\to\overline{C}}\frac{1}{|f|}\sum_{C\in dom(f)}\left[1-BER(C,f(C))\right] \tag{13}$$

$$\max_{f:C\to\overline{C}}\frac{1}{|f|}\sum_{C\in dom(f)}F_1(C,f(C)) \tag{14}$$

where f denotes a correspondence between C and $\overline{C}$. The higher the value of Eq. (13) is, the better does C align to ground-truth circle $\overline{C}$ in terms of BER metric, and the same holds F1-score metric.

4.3. Performance evaluation with different algorithms in literature

Since social circle identifying on ego networks can fall into the domain of community detection [12], hence, in order to answer the first question, we compare the proposed framework DPSCD with following well-known community detection methods which are all open source on our proposed features to answer the first question.

(1) COPRA [39]. Based on the label propagation technique of Raghavan et al. [40], COPRA extends the label and propagation step to include information about more than one community (each vertex can now belong to up to v communities, where v is a parameter which controls the potential degree of overlap between communities) so as to detect overlapping communities in networks.

(2) CONCLUDE [41]. CONCLUDE (COmplex Network CLUster DEtection) is a fast community detection algorithm which consists of three steps: first, (re)weight edges by using a particular random walker; secondly, calculate the distance between each pair of connected nodes; thirdly, partition the network into communities so to optimize the weighted network modularity.

(3) DCM [42]. DCM is an effective algorithm which is able to build well-described cohesive communities starting from any given description or seed set of nodes. It alternates between two phases: a hill-climbing phase producing (possibly overlapping) communities, and a description induction phase which uses techniques from supervised pattern set mining. However, DCM only considers undirected graphs.

(4) CLUTO [43]. CLUTO is a novel clustering framework for multi-topic documents that works as follows: first, each document

Table 1
Statistics of the dataset.

	Facebook	Google+	Twitter
# of nodes	4039	107,614	81,306
# of edges	88,234	13,673,453	1,768,149
# of ego-networks	10	133	1000
# of ground-truth circles	193	479	4869

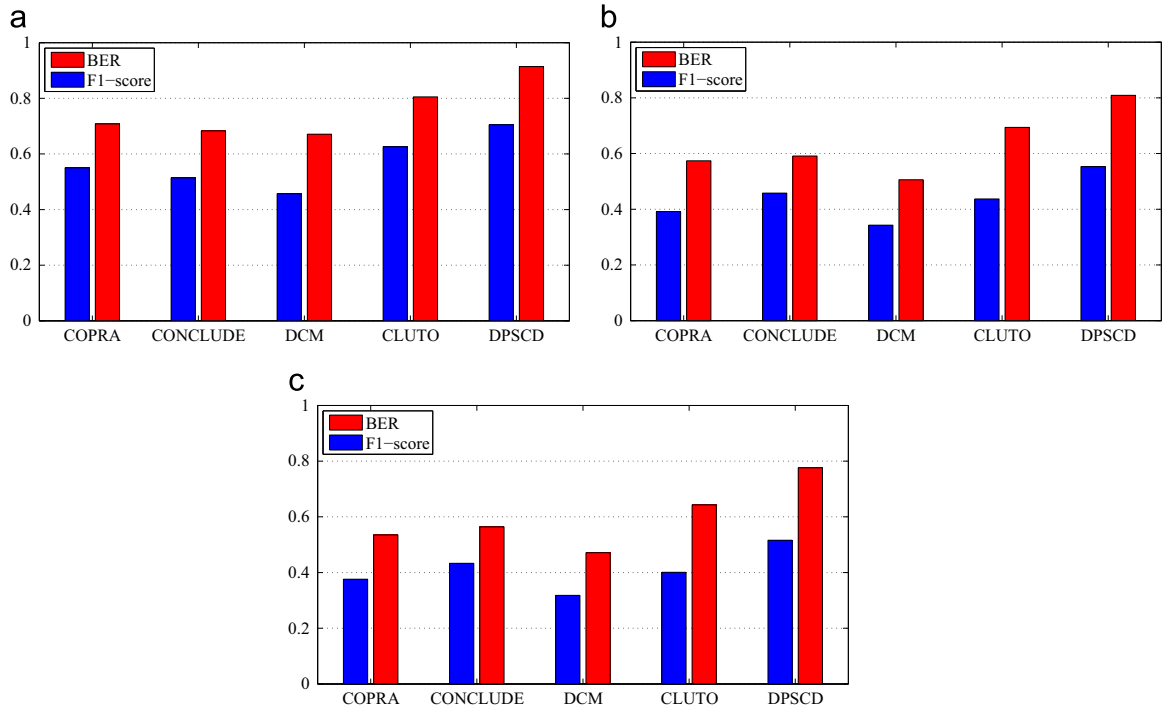


Fig. 1. Performance of different open source community detection methods. (a) Facebook. (b) Google+. (c) Twitter.

in the collection is modeled with a set of segment-sets, which are identified according to the underlying multiple topics of the document; secondly, the segment-sets from all documents are clustered using a document clustering algorithm; thirdly, a possibly "soft" (overlapping) classification of the original documents is induced from the segment-set clustering. Additionally, CLUTO is well-suited for clustering data sets arising in many diverse application areas including information retrieval, customer purchasing transactions, web, GIS, science, and biology.

The BER and F1-score of the methods above are reported in Fig. 1.

It can be observed from Fig. 1(a), (b) and (c), our framework gives better BER and F1-score than other approaches for all datasets. Moreover, results seem to show an enhancement tendency for all approaches when experiment on Facebook. This phenomenon might be explained that Facebook is a network which is constructed based on relationships between acquaintances, so there will be more interactions between users, while in Twitter, users can pay more attention to strangers no matter how many times they interact with each other.

Besides, we further evaluate execution time of different methods along with our method. Fig. 2 shows the execution time, in seconds, used by each method for social circle discovery.

From Figs. 1 and 2, DCM costs the most running time in all cases, while COPRA obtains unsatisfactory results with the smallest computational cost. It can be found that the proposed approach shows the best results with a shorter execution time. Additionally, on the basis of the same original dataset, by employing an improved density peaks-based clustering method to discover social circles with overlap in social networks, we can obtain a significant improvement on performance (+8% in terms of average BER, +15.4% in terms of average F1-score) compared with [9] and achieve better results (+7% in terms of average BER, +2.4% in terms of average F1-score) compared with [12]. The results of aforementioned studies and our proposed method are depicted in Table 2.

In summary, DPSCD can achieve better results even in a small data set with large statistical errors, such as Facebook data set.

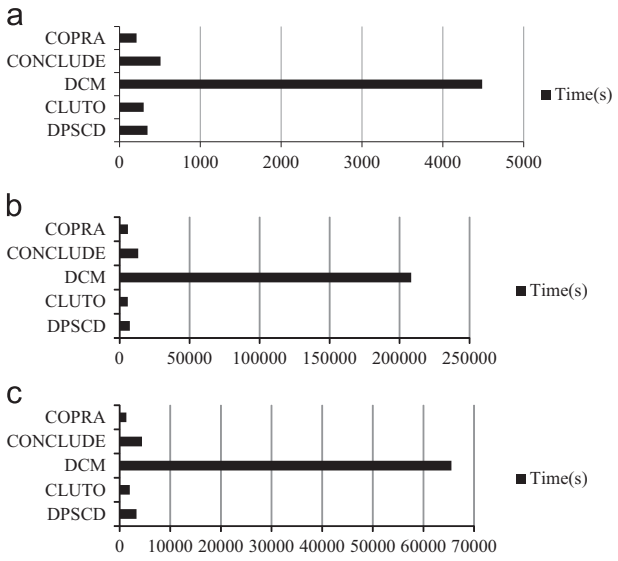


Fig. 2. Execution time, in seconds, of different open source community detection methods. (a) Facebook. (b) Google+. (c) Twitter.

Additionally, compared with other methods, DPSCD can fast detect overlapping social circles accurately through conducting integration on social circles which are clustered via Rodriguez and Laio's method. It follows that all improvement is significant, the proposed framework, DPSCD, outperforms other methods, which answers the first question.

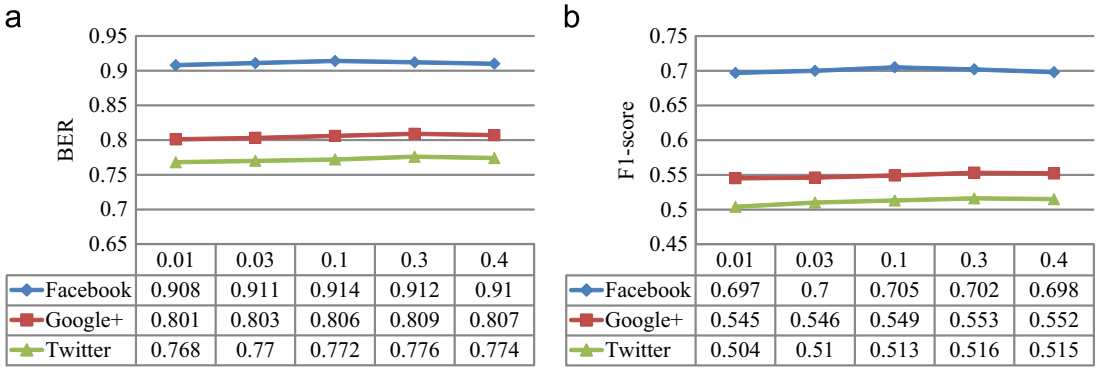
4.4. Analysis on effects of different parameters and different features

4.4.1. Impact of different cutoff distances for social circle discovery

In this section, we use $\sigma = 0.25$, $\alpha = 0.3$ for Facebook and $\alpha = 0.4$ for Google+ and Twitter in all experiments. Firstly, we investigate how much impact that different cutoff distances have

Table 2
The comparison on other social circle discovery works.

Related Works	Facebook		Google+		Twitter		Average	
	BER	F1-score	BER	F1-score	BER	F1-score	BER	F1-score
McAuley [9]	0.84	0.59	0.72	0.38	0.7	0.34	0.753	0.437
Hu and Yang [12]	NA	NA	NA	NA	NA	NA	0.763	0.567
DPSCD	0.914	0.705	0.809	0.553	0.776	0.516	0.833	0.591



	0.01	0.03	0.1	0.3	0.4
Facebook	0.908	0.911	0.914	0.912	0.91
Google+	0.801	0.803	0.806	0.809	0.807
Twitter	0.768	0.77	0.772	0.776	0.774

	0.01	0.03	0.1	0.3	0.4
Facebook	0.697	0.7	0.705	0.702	0.698
Google+	0.545	0.546	0.549	0.553	0.552
Twitter	0.504	0.51	0.513	0.516	0.515

Fig. 3. The impact of cutoff distance d_c in the proposed framework DPSCD. (a) BER. (b) F1-score.

on the performance of our method by varying d_c as $\{0.01, 0.03, 0.1, 0.3, 0.4\}$, BER and F1-score of different values of d_c are reported in Fig. 3.

We draw following observation: being in accord with the foundation in [36], the algorithm is not sensitive to the relative magnitude of d_c in different points which implies that the results are robust with respect to the choice of d_c.

4.4.2. Impact of α for social circle discovery

In this section, we also use $\sigma = 0.25$, $d_c = 0.1$ for Facebook and $d_c = 0.3$ for Google+ and Twitter in all experiments. By changing the parameter α, we explore how different features affect the performance of our method in terms of BER and F1 score. In this paper, α is varied as $\{0, 0.1, 0.3, 0.4, 0.7, 1\}$ and the results are shown in Fig. 4.

It can be observed from Fig. 4, it is difficult to discover social circles with its specific types of features only. Setting very large importance to topological structure-based features and very small importance to profile-based features may result in the worst results. It can be explained that content information encoded in nodes or edges is the essential motivation to attract users to form communities in content-based social networks [44]. In addition, DPSCD achieves the best results with $\alpha = 0.3$ in Facebook and $\alpha = 0.4$ in Google+ and Twitter from which we find that topological structure plays a more important role in Facebook than that in Google+ and Twitter.

The results in Sections 4.4.1 and 4.4.2 further demonstrate the effects of different parameters in DPSCD, which correspondingly answers the second question.

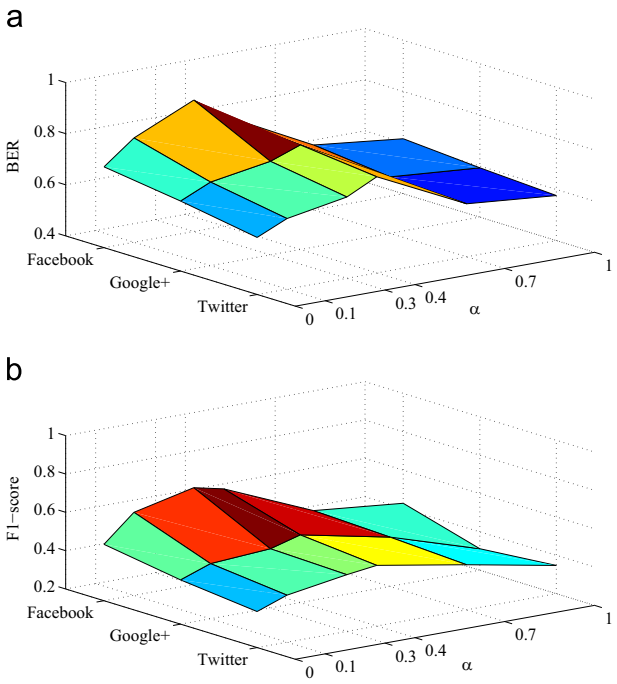


Fig. 4. The impact of parameter α in the proposed framework DPSCD. (a) BER. (b) F1-score.

4.5. Analysis on effects of different features

In this section, we carry on "leave-one-feature-out" experiments on our proposed features to explore the effects that different features have on DPSCD. Due to space restrictions, only average F1-scores are reported in Table 3.

Where hometowns, birthdays, colleagues, political affiliations, education_classes, education_concentration, education_degree, education_school, education_type, education_with, education_year, first_name, gender, languages, last_name, locale, location, work_employer, work_end_date, work_ location, work_position, work_start_date, work_with, work_projects and middle_name are profile-based features in Facebook dataset, gender, last_name, job titles, institutions, universities and places lived are profile-based features in Google+dataset, the set of hashtags and mentions are profile-based features in Twitter dataset.

We draw following observation from Table 3: removing first_name, middle_name, last_name or gender feature lowers the model's prediction abilities, although prediction quality remains relatively high. While removing education-, location-, work- or interests-related

Table 3
Average F1-scores when leaving out different features.

Features	F1-score	Features	F1-score
hometowns	0.570	work_employer	0.581
birthdays	0.590	work_end_date	0.589
colleagues	0.571	work_ location	0.586
political affiliations	0.565	work_position	0.582
education_classes	0.578	work_start_date	0.584
education_concentration	0.573	work_with	0.565
education_degree	0.579	work_projects	0.564
education_school	0.583	middle_name	0.591
education_type	0.588	job titles	0.583
education_with	0.586	institutions	0.584
education_year	0.581	universities	0.588
first_name	0.587	places lived	0.585
gender	0.591	the set of hashtags	0.570
languages	0.567	the set of mentions	0.555
last_name	0.589		
locale	0.588		
location	0.589		

features creates a bigger drop in performance. It can be interpreted that education-, location-, work- and interests-related features are all closely related to user's social life, so they play more important roles in social circle discovery than first_name, middle_name, last_name or gender feature. In conclusion, different features play different roles in social circles discovery, removing either feature may degrade model's performance, our proposed framework that fuse multidimensional features to discover social circles improves the accuracy of social circle discovery algorithm effectively.

In addition, we adopt Salton metric, Jaccard metric and Preferential Attachment metric as topological structure-based feature respectively to evaluate the effect of Salton metric in DPSCD. Jaccard metric supposed that similarity between users was proportional to the ratio of the number of their common neighbors and the number of all their neighbors. In order to measure user's network topological structure in directed social networks, we propose in-link Jaccard metric and out-link Jaccard metric which are shown as follows:

$$inJa(i,v)=\frac{|\Gamma^{in}(i)\cap\Gamma^{in}(v)|}{|\Gamma^{in}(i)\cup\Gamma^{in}(v)|} \tag{15}$$

$$outJa(i,v)=\frac{|\Gamma^{out}(i)\cap\Gamma^{out}(v)|}{|\Gamma^{out}(i)\cup\Gamma^{out}(v)|} \tag{16}$$

where $inJa(i,v)$ and $outJa(i,v)$ stand for in-link Jaccard metric and out-link Jaccard metric between user i and user v respectively. For undirected social networks, we employ traditional Jaccard metric to measure user's network topological structure:

$$Ja(i,v)=\frac{|\Gamma(i)\cap\Gamma(v)|}{|\Gamma(i)\cup\Gamma(v)|} \tag{17}$$

where $Ja(i,v)$ presents Jaccard metric between user i and user v.

Preferential Attachment metric made an assumption that the greater degree user had, the greater possibility user had to establish links with other users. Similarly, we also propose in-link Preferential Attachment metric and out-link Preferential Attachment metric which are shown as follows:

$$inPa(i,v)=|\Gamma^{in}(i)|\times|\Gamma^{in}(v)| \tag{18}$$

$$outPa(i,v)=|\Gamma^{out}(i)|\times|\Gamma^{out}(v)| \tag{19}$$

where $inPa(i,v)$ and $outPa(i,v)$ stand for in-link Preferential Attachment metric and out-link Preferential Attachment metric between user i and user v respectively. For undirected social networks, we employ traditional Preferential Attachment metric to measure user's network topological structure:

Table 4
BERs and F1-scores of different topological structure-based features.

	Facebook		Google+		Twitter	
	BER	F1-score	BER	F1-score	BER	F1-score
Salton metric	0.914	0.705	0.809	0.553	0.776	0.516
Jaccard metric	0.806	0.623	0.711	0.498	0.635	0.427
Preferential Attachment metric	0.754	0.518	0.670	0.432	0.589	0.343

$$Pa(i,v)=\Gamma(i)\times\Gamma(v) \tag{20}$$

where $Pa(i,v)$ presents Preferential Attachment metric between user i and user v.

The BERs and F1-scores of different topological structure-based features are shown in Table 4.

From Table 4, Jaccard metric only takes user's neighbors' set into consideration, Preferential Attachment metric only focuses on the number of user's neighbors, while Salton metric considers both of them, therefore, Salton metric can achieve better results than Jaccard metric and Preferential Attachment metric in all cases. Besides, employing in-link Salton metric and out-link Salton metric instead of traditional Salton metric to measure user's network topological structure in directed social networks can obtain a significant improvement on performance (+23.7% in terms of average BER on Google+ and Twitter, +18.5% in terms of average F1-score on Google+ and Twitter).

The results in Section 4.5 further demonstrate the effects of our proposed features in DPSCD, which correspondingly answers the second question.

5. Conclusion

In this paper, we explored the problem of finding the possible variations and discovering social circles with overlap in social networks. Firstly, in-link Salton metric and out-link Salton metric were presented to measure user's topological structure in directed social networks. And then, given both users' structural and attribute information, an improved density peaks-based clustering method was designed for overlapping social circle discovery. Finally, we ran a set of experiments on three real-world datasets to investigate the performance of our model, and reported system performances in terms of Balanced Error Rate and F1-score. In general, DPSCD approached to the optimum of other social circle discovery algorithms.

In future work, we will speculate on what directions can be undertaken to fuse the information in heterogeneous networks more reasonably. Furthermore, we will employ distributed technologies, such as MapReduce to improve the performance of our method, as well as increase its online application scope.

References

[1] S.H. Li, H. Lou, W. Jiang, J. Tang, Detecting community structure via synchronous label propagation, Neurocomputing 151 (2015) 1063–1075.
[2] L. Wang, J. Wang, Y.J. Bi, W.L. Wu, W. Xu, B. Lian, Noise-tolerance community detection and evolution in dynamic social networks, J. Comb. Optim. 28 (3) (2014) 600–612.
[3] D. Schall, Who to follow recommendation in large-scale online development communities, Inf. Softw. Technol. 56 (12) (2013) 1543–1555.
[4] W.Y. Wang, Y.C. Jiang, Community-aware task allocation for social networked multi agent systems, IEEE Trans. Cybern. 44 (9) (2014) 1529–1543.
[5] B. Yang, X.H. Zhao, J. Huang, D.Y. Liu, Community detection for proximity alignment, Integr. Comput. Aided Eng. 21 (1) (2014) 59–76.

[6] Y.C. Chen, W.Y. Zhu, W.C. Peng, W.C. Lee, S.Y. Lee, CIM: Community-based influence maximization in social networks, ACM Trans. Intell. Syst. Technol. 5 (2) (2014) 25.
[7] Y. Zhang, M. Chen, S.W. Mao, L. Hu, V.C.M. Leung, CAP: community activity prediction based on big data analysis, IEEE Netw. 28 (4) (2014) 52–57.
[8] R.Y.K. Lau, Y.Q. Xia, Y.M. Ye, A probabilistic generative model for mining cybercriminal networks from online social media, IEEE Comput. Intell. Mag. 9 (1) (2014) 31–43.
[9] J. McAuley, Learning to discover social circles in ego networks, in: Proceedings of the 26th Annual Conference on Neural Information Processing Systems, 2012, pp. 548–556.
[10] Y.Y. Ahn, J.P. Bagrow, S. Lehmann, Link communities reveal multi scale complexity in networks, Nature 466 (7307) (2010) 761–764.
[11] M. Girvan, M.E.J. Newman, Community structure in social and biol ogical networks, in: Proceedings of the National Academy of Sciences, 2002, pp. 7821–7826.
[12] Y.M. Hu, B. Yang, Enhanced link clustering with observations on ground truth to discover social circles, Knowl. Based Syst. 73 (2015) 227–235.
[13] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 292–313.
[14] M. Sales-Pardo, R. Guimera, A.A. Moreira, L.A.N. Amaral, Extracting the hierarchical organization of complex systems, Proc. Natl. Acad. Sci. (2007) 15224–15229.
[15] L. Chen, Q. Yu, B.L. Chen, Anti-modularity and anti-community detecting in complex networks, Inf. Sci. 275 (2014) 293–313.
[16] Y.D. Li, J. Liu, C.L. Liu, A comparative analysis of evolutionary and memetic algorithms for community detection from signed social networks, Soft Comput. 18 (2) (2014) 329–348.
[17] L.H. Ma, H. Huang, Q.M. He, K. Chiew, Z.G. Liu, Toward seed-insensitive solutions to local community detection, J. Intell. Inf. Syst. 43 (1) (2014) 183–203.
[18] Q. Cai, M.G. Gong, B. Shen, L.J. Ma, L.C. Jiao, Discrete particle swarm optimization for identifying community structures in signed social networks, Neural Netw. 58 (2014) 4–13.
[19] G.P. Li, Z.S. Pan, B. Xiao, L.W. Huang, Community discovery and importance analysis in social network, Intell. Data Anal. 18 (3) (2014) 495–510.
[20] D. Rhouma, L.B. Romdhane, An efficient algorithm for community mining with overlap in social networks, Expert Syst. Appl. 41 (9) (2014) 4309–4321.
[21] J.T. Qiu, Z.X. Lin, D-HOCS: an algorithm for discovering the hierarchical overlapping community structure of a social network, J. Intell. Inf. Syst. 42 (3) (2014) 353–370.
[22] L. Huang, R.X. Li, H. Chen, X.W. Gu, K.M. Wen, Y.H. Li, Detecting network communities using regularized spectral clustering algorithm, Artif. Intell. Rev. 41 (4) (2014) 579–594.
[23] Z. Wu, J. Cao, J.J. Wu, Y.Q. Wang, C.Y. Liu, Detecting genuine communities from large-scale social networks: a pattern-based method, Comput. J. 57 (9) (2014) 1343–1357.
[24] T.A. Dang, E. Viennet, Community detection based on structural and attribute similarities, in: Proceedings of the Sixth International Conference on Digital Society, 2012, pp.7–12.
[25] J.T. Sang, C.S. Xu, Right buddy makes the difference: an early exploration of social relation analysis in multimedia applications, in: Proceedings of the 20th ACM Multimedia Conference, 2012, pp.19–28.
[26] M. Yan, J.T. Sang, T. Mei, C.S. Xu, Friend transfer: cold-start friend recommendation with cross-platform transfer learning of social knowledge, in: Proceedings of the 2013 IEEE International Conference on Multimedia and Expo, 2013, pp. 1–6.
[27] Z. Yin, L. Cao, Q. Gu, J. Han, Latent community topic analysis: integration of community discovery with topic modeling, ACM Trans. Intell. Syst. Technol. 3 (4) (2012) 67–83.
[28] J.F. Zhuang, T. Mei, S.C.H. Hoi, X.S. Hua, Y.D. Zhang, Community discovery from social media by low-rank matrix recovery, ACM Trans. Intell. Syst. Technol. 5 (4) (2015) 1–19.
[29] Z. Wang, D.Q. Zhang, X.S. Zhou, D.Q. Yang, Z.Y. Yu, Z.W. Yu, Discovering and profiling overlapping communities in location-based social networks, IEEE Trans. Syst. Man Cybern.: Syst. 44 (4) (2014) 499–509.
[30] X. Yu, J. Yang, Z.Q. Xie, A semantic overlapping community detection algorithm based on field sampling, Expert Syst. Appl. 42 (1) (2015) 366–375.
[31] D. William, W. Hu, Mutually enhancing community detection and sentiment analysis on twitter networks, J. Data Anal. Inf. Process. 3 (2013) 19–29.
[32] A. Mislove, B. Viswanath, P.K. Gummadi, P. Druschel, You are who you know: inferring user profiles in online social networks, in: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, 2010, pp. 251–260.
[33] A. Silva, W. Meira, M.J. Zaki, Mining attribute-structure correlated patterns in large attributed graphs, in: Proceedings of the VLDB Endowment, 2012, pp. 466–477.
[34] M.E. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2) (2001) 025102.
[35] Z.W. Yu, X.S. Zhou, D.Q. Zhang, G. Schiele, C. Becker, Understanding social relationship evolution by using real-world sensing data, World Wide Web 16 (5–6) (2013) 749–762.
[36] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.
[37] A. Lancichinetti, S. Fortunato, J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks, New J. Phys. 15 (2009).
[38] J. Yang, J. Leskovec, Defining and evaluating network communities based on ground-truth, Knowl. Inf. Syst. 42 (1) (2015) 181–213.
[49] S. Gregory, Finding overlapping communities in networks by label propagation, New J. Phys. 12 (10) (2010) 2011–2024.
[40] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E: Stat. Nonlinear Soft Matter Phys. 76 (3) (2007) 036106.
[41] P. DeMeo, E. Ferrara, G. Fiumara, A. Provetti, Mixing local and global information for community detection in large networks, J. Comput. Syst. Sci. 80 (1) (2013) 72–87.
[42] S. Pool, F. Bonchi, M.V. Leeuwen, Description-driven community detection, ACM Trans. Intell. Syst. Technol. 5 (2) (2014) 1–28.
[43] A. Tagarelli, G. Karypis, A segment-based approach to clustering multitopic documents, Knowl. Inf. Syst. 34 (3) (2013) 563–595.
[44] C.D. Wang, J.H. Lai, P.S. Yu, NEIWalk: community discovery in dynamic content-based networks, IEEE Trans. Knowl. Data Eng. 26 (7) (2014) 1734–1748.

Mengmeng Wang, born in 1987. Ph.D. candidate in the College of Computer Science and Technology, Jilin University, Changchun, China. Her main research interests include Web information retrieval and mining, ontology, Deep Web, machine learning and social network analysis.

Wanli Zuo, born in 1957. Professor and Ph.D. supervisor in the College of Computer Science and Technology, Jilin University, Changchun, China. Senior member of China Computer Federation. His main research interests include Web information retrieval and mining, ontology, Deep Web, machine learning and social network analysis.

Ying Wang, born in 1981. Received her Ph.D. degree in the College of Computer Science and Technology from Jilin University in 2010. Member of China Computer Federation. Currently she is lecturer in the College of Computer Science and Technology, Jilin University, Changchun, China. Her main research interests include Web information retrieval and mining, ontology, Deep Web, machine learning and social network analysis.