Overlapping Community Structure Detection in Multi-Online Social Networks

Wei Fan*, Kai-Hau Yeung*, Wenjie Fan†
*Department of Electronic Engineering, City University of Hong Kong, Hong Kong
†Department of Information Systems, City University of Hong Kong, Hong Kong

***Abstract*—Overlapping community structures can reflect the fact that an individual can belong to more than one community, especially the individuals in social networks. It is important to detect overlapping community structures in online social networks. In this paper, an overlapping community detection algorithm is proposed for online social networks. Different from previous works, this algorithm can be applied in both undirected and directed networks. Moreover, for people who have accounts in several online social networks, this algorithm can be used to combine community structures of different networks. We assume that users' community structure in an online social network is actually a reflection of users' community structure in real life. By combining different reflections with the proposed algorithm, detected community structure can match users' interest-based communities better.**

***Keywords*-Online social networks; Overlapping community structure**

I. Introduction

In recent years, online social networking service has become popular. An online social networking service is a popular web-based service which provides a platform for people to create profiles and share these profiles with other people. With this service, people can share their daily life and their opinions with their friends on the Internet. Some online social networking sites such as MySpace, Facebook, Google+, and Bebo are popular and have attracted millions of users [1], [2]. Some people even register on several online social networking sites. In this work, we focus on a set of users who have accounts of three online social networking sites, namely Facebook, Foursquare and Twitter. For each website, users and the connections between them can constitute a network. We call the networks of these online social networking sites the "online social networks".

Community structure is an important feature of networks because nodes in the same community often share common properties. Community structure of online social networks helps people understand users' behavior and interests in these networks better. Moreover, this structure helps social networking service providers offer better services and recommendations for their users. Researchers have tried to study the community structure of online social networks and investigate the common characteristics of members in detected communities. Community structure has been detected in many online social networks, such as Twitter, Delicious, Facebook and so on [3]–[7].

Researchers have proposed methods to detect communities in networks based on networks' connectivity structures. There are two types of community detection methods: non-overlapping methods and overlapping methods. For example, the methods proposed in [8]–[10] are non-overlapping community detection methods. Community structures detected by these methods are called "partitions". A node in a partition can be assigned to only one community. Community structures detected by overlapping methods are called "covers" by the authors of [11]. In social networks, people can be members of more than one cluster in most cases. For example, people can be in a community of their family while they can at the same time be in a community of their college friends. Overlapping community structures can reflect this characteristic.

There have been works on overlapping community structure detection. In [12], the authors provide an extension of modularity for overlapping community structure. And this extension can be used in directed networks. But their definition has a limitation. Suppose the probability that a node i belongs to community c is $a_{i,c}$, their definition requires that $\sum_{c \in C} a_{i,c} = 1$ [13], where C is the set of communities. This means that if node i belongs to a community strongly, this node cannot belong to another community strongly at the same time. This is different from the real social networks. In [14], [15], overlapping community detection methods which do not have this limitation are proposed. But methods in these two works do not consider directed networks. Therefore, we propose an algorithm which can detect overlapping communities in both directed and undirected networks. This algorithm will allow nodes to belong to different communities fully. And we can combine the community structures of different online social networks easily with this algorithm. To evaluate our algorithm, we will measure the detected community structure by comparing it with users' interest-based communities.

II. Dataset

In this paper, three websites which provide online social networking services are studied: Facebook, Twitter and Foursquare. These websites are popular and have attracted millions of users. Although all these three websites are providing online social networking services, their focuses are different. Facebook provides general services: photos/videos uploading, blog sharing, and applications. But Twitter and

978-1-4799-1866-9/15/$31.00 ©2015 IEEE

Foursquare offer simpler services than Facebook does. Twitter is a micro-blogging website allowing users to post updates which should be within 140 letters. People can share their ideas or opinions quickly on Twitter. And Foursquare is a location-based social networking website. Users of this website can check in at some venues with computers or with their cell phones which enable global-positioning-system(GPS)-related applications. They can share their current locations and some information of these venues with their friends. A set of people who have accounts in these three online social networks are studied. There are 270 users in our dataset. These users and the friendships between them in different online social networks can construct three networks.

In online social networks, a user can create a list of other users and share the list on his/her profile to indicate their friendship. In this paper, we mainly focus on the friendship of users in online social networks. In Facebook and Foursquare, this friendship is called "friend". Both of these two social networks are undirected networks, which means that if two users are "friends", they will appear in each other's friend list. But in Twitter, there are two kinds of friendship: "following" and "followed". A user can track some users that he/she is interested in, which means "following". And this user can also be tracked, or "followed" by other users. So the network of Twitter is a directed one. In this paper, a network can be presented as an adjacency matrix $\boldsymbol{A}$. A_{ij} is the weight of the edge between nodes i and j. In unweighted and undirected networks, $A_{ij} = 1$ if there is an edge connecting these two nodes, and $A_{ii} = 0$. In directed networks, $A_{ij} = 1$ if there is an edge from node i to node j. Adjacency matrices of these three networks can be labeled $\boldsymbol{A}_{fsq}$, $\boldsymbol{A}_{fb}$ and $\boldsymbol{A}_{tw}$.

III. The proposed algorithm

Many quality functions have been defined to describe characteristics of communities, such as modularity [16] and internal density, but "no definition is universally accepted" [11]. The most commonly accepted characteristic of communities is that communities should be dense [14]. This means that edges within a community should be more than the edges connecting this community and the rest of the network. So our algorithm is based on the density of communities.

In [15], it is mentioned that in overlapping community structures, communities are local and independent of their topological environment. And the authors provide an algorithm to classify nodes into local communities. In this paper, the first stage of our algorithm is to detect local communities. For each node in a network, we first group its neighbors and itself into a local community c. For each node i in this community c, we calculate the difference between its degree going within c and its degree going out of c:

$$x_i = \sum_{j \in c, i \neq j} a_{ij} - \sum_{j \notin c} a_{ij} \tag{1}$$

Here $a_{ij} = 1$ if nodes i and j are connected, and $a_{ij} = 0$ otherwise. Then we can obtain the sum of the differences $f = \sum_{i \in c} x_i$ across the initial local community. $f > 0$ means that the edges in this community are more than the edges connecting this community and the rest of the network. If $f \leq 0$, we will remove the node with the lowest x_i and calculate x_i for all remaining nodes in community c again. This step is repeated until $f > 0$. Note that our algorithm requires that the sizes of local communities should be larger than 1. For directed network (Twitter), we will try to construct two local communities for each node based on its out degree and in degree, respectively. So we have two equations to calculate x_i:

$$x_i^{out} = \sum_{j \in c, i \neq j} a_{ij} - \sum_{j \notin c,} a_{ij} \tag{2}$$

$$x_i^{in} = \sum_{j \in c, i \neq j} a_{ji} - \sum_{j \notin c,} a_{ji} \tag{3}$$

The steps to detect local communities are shown in Algorithm 1. The complexity of Stage 1 is $O(N)$, where N is the number of nodes in a network.

Algorithm 1 Stage 1

Input: Friendship networks $\boldsymbol{A}$;

$C = \emptyset$;
for each node i in $\boldsymbol{A}$ **do**
  find the set of friends of i: $Neighbor(i)$;
  $c_{temp} = \{i\} \cup Neighbor(i)$;
  $f = 0$;
  while $f \leq 0$ **do**
    for each node i in c_{temp} **do**
      calculate the value of x_i;
    end for
    $f = \sum_{i \in c} x_i$;
    if $f \leq 0$ **then**
      remove the node with the lowest x_i from c_{temp};
    end if
  end while
  if $|c_{temp}| > 1$ **then**
    $C = [C, c_{temp}]$;
  end if
end for;

After the first stage of our algorithm, we can obtain a set of local communities. These local communities overlap a lot. Some local communities even have many common nodes. In [14], [15], the authors discuss when to separate two overlapped communities and when to combine them. In our algorithm, we follow the second phase of the algorithm

in [15]: to merge two local communities if their overlapping score is larger than a parameter. In [15], the overlapping score considers both common nodes and common edges. But in our algorithm, we only take common nodes into account so that we can combine local communities of different online social networks more easily. In our algorithm, if two communities c_i and c_j have an overlapping score $(|c_i \cap c_j|)/\min\{|c_i|, |c_j|\} > \beta$ $(0 < \beta < 1)$, we merge these two communities, replace the community with a smaller index with the merger, and delete the community with the larger index. In the second stage of our algorithm, each pair of communities is checked. And then communities are merged until no more communities can be merged. The complexity of Stage 2 is $O(N^2)$.

When we detect community structure of one single online social network, we can first detect local communities in Stage 1. Then we merge these detected local communities in Stage 2 of our algorithm. For example, we can obtain three set of communities C_{fsq}^{ini}, C_{fb}^{ini}, and C_{tw}^{ini} (from Foursquare, Facebook and Twitter, respectively)after Stage 1. For these three online social networks, we can obtain C_{fsq}, C_{fb}, and C_{tw} after Stage 2.

When we combine community structures of several online social networks, we first let $C_{all}^{ini} = C_{fsq}^{ini} \cup C_{fb}^{ini} \cup C_{tw}^{ini}$. C_{all}^{ini} is the input of the Stage 2. Then we can obtain C_{all}.

IV. Evaluation

The proposed algorithm is applied in our dataset. Four sets of communities are obtained: C_{fsq}, C_{fb}, C_{tw} and C_{all}. In Figure 1, we show how the number of detected communities in these four sets changes as the value of β changes. The number of detected communities in C_{all} is close to the one in C_{tw} when $\beta \leq 0.85$. If we have $\beta = 0.9$, this value is too high so that many local communities cannot be combined. So the number of communities in C_{all} rises a lot .

In Figure 2, we plot the correlation between the size of detected communities and β. The case in C_{all} is similar to the one in C_{tw} again. And the size of communities in C_{all} is the largest. It may be because there are more local communities that can be combined in C_{all}^{ini}.

Then we evaluate the proposed algorithm by comparing detected communities and users' interest-based communities. Interest-based communities are used to evaluate the performance of different community definitions in [17]. In our work, interest-based communities are constructed with the information of users' favorite pages. "Like" is one of Facebook's most well-known features. If a user thinks the content on a page is interesting, he/she can click the button "like" on this page. And users can "like" the pages of their colleges, favorite movies, and so on. These pages are also called "favorite pages" in this paper. Users who like the same page can construct an interest-based community. We delete the interest-based communities which have only one member. The set of interest-based communities is denoted

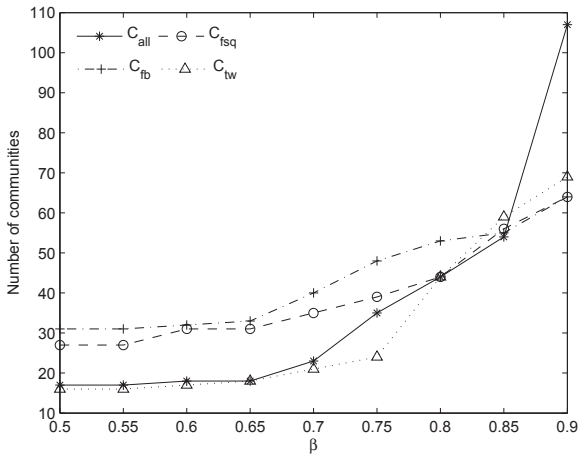

Figure 1. Correlation between the number of detected communities and β

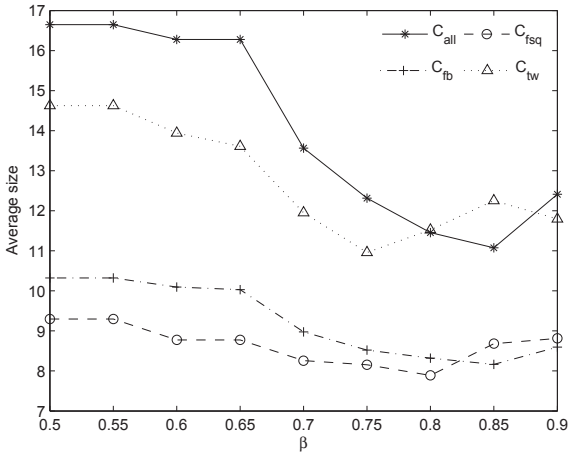

Figure 2. Correlation between size of detected communities and β

by C_{like}. We suppose that these interest-based communities are "real" communities of users.

There are many methods to compare two sets of communities. For example, in [18], the authors compared some community detection methods based on computer-generated networks with known community structures. They considered fraction of nodes correctly identified. But interest-based communities are numerous and overlap a lot. We cannot calculate the fraction of nodes correctly identified. So we compare the set of detected communities C to the set of communities constructed with like information C' using the evaluation function in [19]:

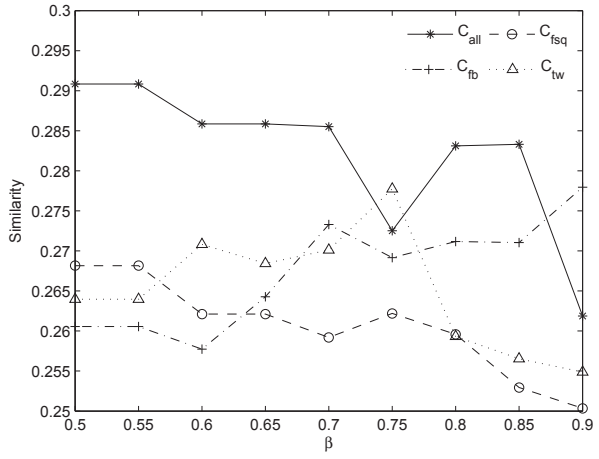

Figure 3. Comparison of detected communities and interest-based communities

$$S = \frac{1}{2|C'|}\sum_{c'_i \in C'} \max_{c_j \in C} \delta(c'_i, c_j) + \frac{1}{2|C|}\sum_{c_i \in C} \max_{c'_j \in C'} \delta(c'_i, c_j) \tag{4}$$

Here $\delta(c'_i, c_j)$ is Jaccard similarity between communities c'_i and c_j:

$$J(c'_i, c_j) = \frac{|c'_i \cap c_j|}{|c'_i \cup c_j|} \tag{5}$$

The value of S is between 0 and 1. And larger values indicate that the set of detected communities is more similar to the set of communities constructed with like information.

We compare detected communities with interest-based communities in Figure 3. It is found that when we have $0.5 \leq \beta \leq 0.7$ and $0.8 \leq \beta \leq 0.85$, C_{all} performs the best at matching interest-based communities with values of similarity 0.283–0.291. The reason may be that C_{all} combines the connectivity information of all three online social networks, so that it can reflect the underlying community structure of users best. And when β is small ($0.5 \leq \beta \leq 0.6$), C_{fb} performs worse than the other three community structures. But when we have $0.7 \leq \beta \leq 0.9$, C_{fb} matches C_{like} better than C_{fsq} and C_{tw} in most cases.

V. Conclusion

Community structure of online social networks is crucial to help researchers analyze people's behavior, especially overlapping community structure, which can capture the characteristic that a user can belong to several communities. As online social networking service becomes popular, there are more and more websites built to provide such services. And these online social networks may be quite distinct from each other. For example, Twitter is a directed network while Facebook is an undirected one. Moreover, people may register in more than one online social network and construct different communities. It will be helpful if there is a community detection algorithm which is applicable in different kinds of networks. But existing community detection algorithms are unable to achieve this goal. Existing algorithms cannot deal with the situation that a set of people construct different community structures in different online social networks, neither.

To fill this gap, we have proposed an overlapping community detection algorithm for multi-online social networks in this paper. This algorithm can detect overlapping communities, which can reflect the community structures in real social networks. In the first stage of this algorithm, local communities need to be detected. We constructed the initial local communities with each node and its neighbors. For directed networks, we defined local communities based on nodes' in degrees and out degrees. So this algorithm works in both directed and undirected networks. Considering the situation that people may register in more than one online social network, our algorithm is designed to combine the community structures of different online social networks. In the second stage of our algorithm, some local communities are merged if they overlap a lot. We calculated the overlapping score based on common nodes of two local communities, so that we can combine local communities of different online social networks without knowledge of common edges.

We assume that people's community structure in an online social network is a reflection of the underlying community structure in their real life. With the proposed algorithm, we can combine different reflections to approach the underlying community structure. We used users' interest-based communities to evaluate our detection algorithm. And we applied a method to measure the similarity between detected community structure and interest-based community structure. It is found that in most cases the combination of communities of the three online social networks in our dataset matches interest-based communities better than the community structure of one single online social network does, around 4% to 12% increment of similarity. The average value of similarity is about 0.261–0.267 if we compare the interest-based community structure with the detected community structure of one single online social network. But this value is 0.282 if we combine the communities of three online social networks. So our overlapping community detection algorithm can help us find out the underlying "real" communities. As this algorithm can combine different community structures easily, it also provides a quick way to incorporate different sources of information about community structure.

References

[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated*

Communication, vol. 13, no. 1, pp. 210–230, 2007.

[2] M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang, "Poking facebook: Characterization of osn applications," in *Proceedings of the First Workshop on Online Social Networks*, ser. WOSN '08. New York, NY, USA: ACM, pp. 31–36, 2008.

[3] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, "Comparing community structure to characteristics in online collegiate social networks," *SIAM Review*, vol. 53, no. 3, pp. 526–543, Jan. 2011.

[4] Z. Zhang, Q. Li, D. Zeng, and H. Gao, "User community discovery from multi-relational networks," *Decision Support Systems*, vol. 54, no. 2, pp. 870–879, Jan. 2013.

[5] Y. Zhang, Y. Wu, and Q. Yang, "Community discovery in twitter based on user interests," *Journal of Computational Information Systems*, vol. 8, no. 3, pp. 991–1000, 2012.

[6] E. Ferrara, "A large-scale community structure analysis in facebook," *EPJ Data Science*, vol. 1, no. 1, pp. 1–30, Dec. 2012.

[7] ——, "Community structure discovery in facebook," *International Journal of Social Network Mining*, vol. 1, no. 1, pp. 67–90, Jan. 2012.

[8] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, 066111, Dec. 2004.

[9] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, 066133, Jun. 2004.

[10] ——, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, 036104, Sep. 2006.

[11] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 35, pp. 75–174, Feb. 2010.

[12] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 03, p. P03024, Mar. 2009.

[13] T. Li, J. Liu, and W. E, "Probabilistic framework for network partition," *Physical Review E*, vol. 80, no. 2, 026106, Aug. 2009.

[14] A. Lzr, D. bel, and T. Vicsek, "Modularity measure of networks with overlapping communities," *EPL (Europhysics Letters)*, vol. 90, no. 1, p. 18001, Apr. 2010.

[15] N. Nguyen, T. Dinh, D. Nguyen, and M. Thai, "Overlapping community structures and their detection on social networks," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pp. 35–40, Oct. 2011.

[16] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, 026113, pp. 1–15, Feb. 2004.

[17] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, ser. MDS '12. Beijing, China: ACM, 2012.

[18] L. Danon, A. Daz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 09, p. P09008, Sep. 2005.

[19] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," *arXiv:1401.7267 [physics]*, Jan. 2014.