Community Extraction from Massive Social Networks

Jie Cao, Zhiang Wu
Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics,
caojie690929@163.com; zawuster@gmail.com

Junjie Wu*
*Corresponding author: School of Economics and Management, Beihang University, wujj@buaa.edu.cn

In this paper, we propose a framework named CEF for community extraction from large-scale social networks. The key idea is to isolate the interference of weakly connected nodes by efficiently extracting core nodes based on the novel *structural approximation* concept. Experiments on various real-world social networks validate the advantages of CEF.

Key words: Social Network; Community Detection; Community Extraction; Structural Approximation

1. Introduction

Recent years have witnessed an increasing interest in detecting closely-knit groups in large-scale social networks of various kinds. The groups are also called *communities*, *clusters*, *cohesive subgroups* or *modules* in different research fields [15]. In general, actors in a same community tend to interact with one another more frequently than with those outside the community. Since social networks have been deeply integrated into various business processes, such as network marketing, online recommendation and location-based services, detecting meaningful and useful communities from massive social networks has become a critical preliminary for many subsequent business tasks.

Though being extensively studied [11, 6, 12], community detection remains a core problem in social network analysis. The existing methods in the literature can be roughly divided into two categories, one is with global models and the other is not. The methods with global models include mixture models (or simply K-means model), latent space models [7], stochastic block models, spectral clustering [5], and modularity [11], among others. These methods typically consider the global topology of a network, and aim to optimize a criterion defined over a network partition. These criteria are often designed carefully so that the optimizations could be achieved via some efficient yet robust heuristics. The problem is, a real-life network might probably contain nodes that have weak connections to any communities, as illustrated by the Karate club network case discussed later in this paper. In such cases, the global models typically split up weakly connected nodes and group them together with tighter communities, which actually impedes us from finding the real communities.

The methods without global models typically employ a bottom-up strategy to find communities. They often start by defining the properties of a node, a pair of nodes, or a group of nodes in a same community, and then search within a whole network for the communities that hold the proposed properties [14]. In these studies, a community could be regarded as a clique, a k-clique, a k-club, a quasi-clique [15], an equivalent structure, or the combination of node pairs that have nodes similar to each other, as measured by for example Jaccard coefficient or cosine similarity. These methods attempt to find well-defined and meaningful communities, and might avoid the negative influences from weakly connected nodes. However, they often suffer from the computational inefficiency; that is, the exhaustive search of a defined structure is often prohibitively expensive, especially when facing the big data emerging from ever-growing social media.

In this paper, we propose a novel framework for community detection. The core idea is to find a balance between the above two types of methods so that we can find meaningful communities in an efficient manner. The key features of our framework lie in the following two aspects. First, we look for tight groups that contain nodes in *structural quasi-equivalence*, called *core nodes*. An efficient algorithm is proposed to search for the core nodes, and the weakly connected nodes will be isolated to avoid confusion. Second, our framework can incorporate any existing methods with global models for network partitioning. This guarantees the efficiency and provides flexibility in real-world applications. Due to the above two features, it is more precise to describe our work as "community extraction"; that is, we attempt to extract tight communities with only core nodes from massive social networks.

Finally, we briefly go through few work related to community extraction. Ref. [1] tried to divide nodes into core and periphery sets based on the proposed CP measure, but their methods can only work for small-size networks. Local community detection [3] aims to find the tightest community around a given node locally rather than globally. Recently, Ref. [18] proposed a criterion W to extract tight communities one by one, but the tabu search prevents it from being further used for large-scale networks. Some hierarchical models also seek to highlight communities by excluding unrelated nodes [4].

2. Community Extraction Framework

In this section, we introduce our framework for community extraction. We begin by giving some basic notations. An undirected network is often denoted as $N=(V,E)$, where V is the set of nodes in N, and E is the set of edges that connect the nodes in V. Community extraction is formulated as finding a crisp or overlapped partition $V^c = V_1 \bigcup \cdots \bigcup V_K$, with $V^c \subseteq V$ consisting of the *core nodes*. A network with $|V|=n$ nodes can be also represented by a $n \times n$ adjacency matrix $A=[A_{ij}]$, where $A_{ij}=1$ if there is an edge between nodes i and j and $A_{ij}=0$ otherwise. For undirected networks, A is symmetric obviously.

2.1. Structural Approximation

Intuitively, V_c should be compact enough to contain only core nodes. Moreover, V_c is expected to have an apparent structure so that $V_1, \cdots, V_K$ are well separated. In the literature, a key related concept is *structural equivalence*. That is, two nodes i and j are structurally equivalent, if $\forall\ k \neq i,j,\ e(i,k) \in E$ iff $e(j,k) \in E$. In other words, structurally equivalent nodes must have totally overlapped friends. Following this clue, we could expect that the nodes of the same equivalence class form very tight communities. This idea, however, confront one problem. Structural equivalence is often too restrictive for real-life social networks — we can find very few communities in such equivalence. Other existing definitions of equivalence, such as *automorphic equivalence* and *regular equivalence*, are looser than structural equivalence but suffer from the very high computational cost.

In what follows, we propose a new statistic G to extract from massive social networks the communities in *structural approximation*. Let N_i denote the set of neighbors of node i, i.e., $k \in N_i$ iff $e(k,i) \in E$. Then for any node set S with $|S| \geq 2$, we define

$$G(S,p) = \left(\frac{1}{|S|} \sum_{i=1}^{|S|} \left(\frac{|N_1 \bigcap \cdots \bigcap N_{|S|}|}{|N_i|} \right)^{-p} \right)^{\frac{1}{-p}},\ p \in [0,+\infty). \tag{1}$$

Apparently $G \in [0,1]$. We say S is structurally p-approximate with respect to τ, denoted as $S_\tau^{(p)}$, iff $G(S,p) \geq \tau$, $\tau \in [0,1]$. Intuitively, G measures the ratio of common friends among all the friends of the nodes in a group. A larger G indicates a generally higher percentage of common friends, and the nodes are thus apt to form a tighter community. For the properties of G, we have a proposition as follows.

Proposition 1. *G has the following properties:*
1. *(Similarity). Assume $|S|=2$. Then $G(S,p)$ reduces to the cosine similarity if $p=0$.*
2. *(Asymptotics). $S_\tau^{(p)}$ tends to be structurally equivalent as τ increases to 1.*
3. *(p-Monotonicity). $S_\tau^{(p)}$ is structurally p'-approximate w.r.t. τ if $p' \leq p$.*

The proof is omitted due to the page limit. Properties 1 and 2 reveal the strong correlations between G and the well-known vertex similarity concepts: cosine similarity and structural equivalence. We can therefore expect that a structurally approximate $S_\tau^{(p)}$ could be a tight community that contains very similar members. τ here serves as a valve that controls the number and quality of $S_\tau^{(p)}$'s we can obtain — a larger τ will result in fewer but generally tighter $S_\tau^{(p)}$'s. Property 3 indicates the monotone property of G in p; that is, given τ fixed, a larger p will lead to fewer but generally tighter $S_\tau^{(p)}$'s. By default, we set $p=0$ and leave the valve function to τ.

2.2. Extraction Method

To extract all $S_\tau^{(p)}$ in a brute-force manner is prohibitively expensive. We here propose a method to boost the search process. Let d_i denote the degree of node i, we first have the following theorem:

Theorem 1. *Let $S_+ = S \bigcup i$, $i \notin S$. Then $G(S_+,p) < G(S,p)$ if $d_i > d_j, \forall j \in S$.*

The proof is omitted due to the page limit. Theorem 1 actually implies a stopping criterion for the search of $S_\tau^{(p)}$. That is, if S is not structurally p-approximate w.r.t. τ, then S_+ will definitely not be either, given that S_+ is formed by adding one or more nodes of higher degrees to S. Therefore, the cost for testing potentially tremendous S_+'s can be saved.

To make use of Theorem 1, we should specify an appropriate examination sequence for all the possible node sets. Apparently, this should be a top-down sequence, from small groups to large groups by adding nodes one after another. The nodes should be ranked in the increasing order of degree so that they are added to the groups in strict accordance with this order. Fig. 1 illustrates the sequence by a small network containing only five nodes with $d_A \leq \cdots \leq d_E$. For instance, if $S=\{A,B\}$ is not structurally p-approximate, then all its subsequent node sets $\{A,B,C\}$, $\{A,B,D\}$, $\{A,B,E\}$, $\{A,B,C,D\}$, $\{A,B,C,E\}$, $\{A,B,D,E\}$, and $\{A,B,C,D,E\}$ will not be structurally p-approximate definitely. As a result, we can avoid examining these sets.

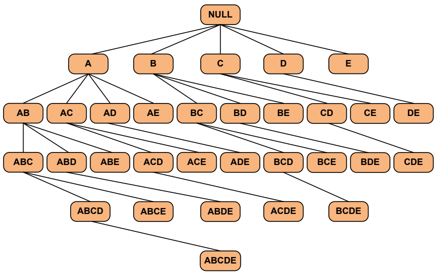


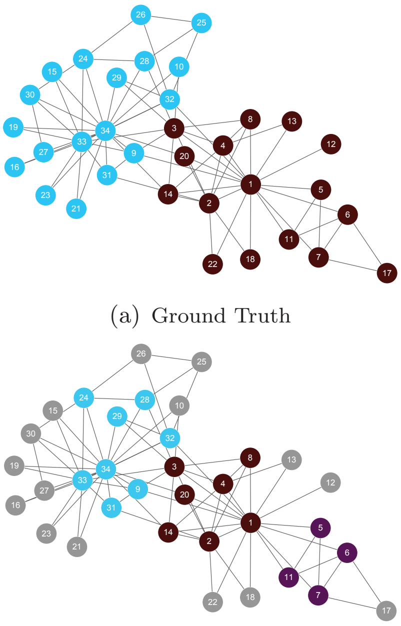
Figure 1 Examination Sequence

We have implemented the above extraction method, named SAM (Structural-Approximation-Mining), that takes p, τ and σ as the major arguments. σ is a threshold used in conjunction with τ to avoid chance communities; that is, for any $S_\tau^{(p)}$, we also require $|N_1 \bigcap \cdots \bigcap N_{|S|}|/|N| \geq \sigma$. Due to the page limit, we will not cover the algorithmic details of SAM any more.

2.3. The Framework

Note that the extracted $S_{\tau}^{(p)}$'s may be overlapped, which are then combined to obtain the set of core points: V^c. So the remaining work is to detect communities from V^c. Since V^c is often far smaller than V, we can simply employ some existing method to fulfill this task. For instance, we can extract the adjacency matrix A^c from A for V^c, and then call K-means clustering on A^c to get a partition of V^c. Or we can extract the subnetwork N^c from N for V^c, and then apply graph partitioning to N^c to get the communities.

Until now, we have the complete framework, named CEF (Community-Extraction-Framework), for community extraction from social networks. CEF consists of three steps. The first step is to extract core nodes from the whole network using SAM, the second step is to partition core points into communities, and the final step is to evaluate the communities. Fig. 2 illustrates the performance of CEF on a famous real-world network: `Karate Club` [17]. As can be seen, compared with the ground truth, CEF does find the core nodes (in non-grey colors) of the two factions, and even identifies a sub-faction (nodes in dark-blue) inside one faction.



(a) Ground Truth

(b) Our Result

Figure 2 Karate Club

3. Experimental Validation

In this section, we demonstrate the effectiveness of CEF for real-world community detection. Three real-life social networks were used here, i.e., `Oklahoma`, `Enron` and `Gowalla`, with some statistics given in Table 1. `Oklahoma` is a Facebook subnetwork representing the friendships in University of Oklahoma [16]. `Gowalla` is also a friend network but built on trajectory information [2]. `Enron` is an email communication network covering half million emails [10].

Table 1 Experimental Data sets.

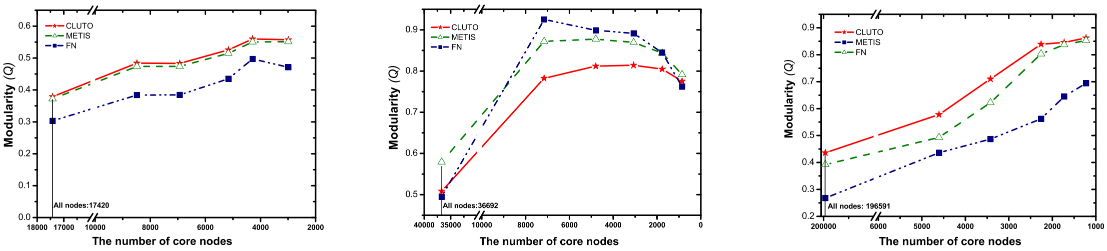
Data set	$\|V\|$	$\|E\|$	$<k>$	C
`Oklahoma`	17420	892524	102.47	0.23
`Enron`	36692	367662	20.04	0.42
`Gowalla`	196591	1900654	19.34	0.20

$<k>$: average degree; C: clustering coefficient.

SAM was employed for the extraction of core nodes with $p=0$. Three tools, i.e., CLUTO, Fast-Newman (FN), and METIS, were adopted to partition the core nodes into communities. CLUTO [8] is a widely used implementation of K-means for text clustering, METIS [9] is a popular graph partitioning tool, and FN [11] is an agglomerative hierarchical clustering method based on the modularity concept.

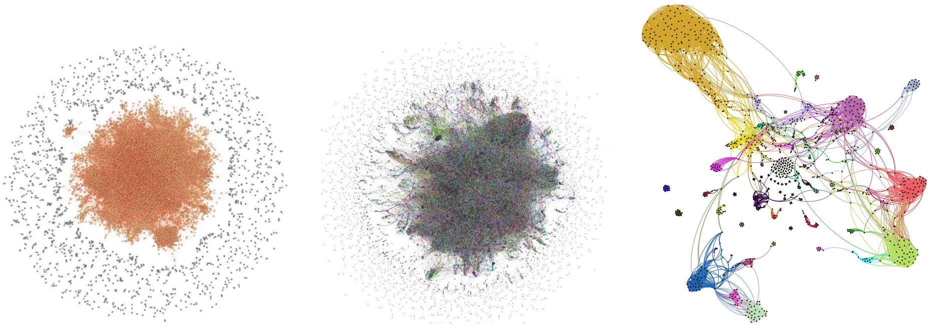
Since we do not have the ground truth or semantic information of these networks, we used the widely adopted Q function [15, 13] as internal measure to evaluate the overall quality of extracted communities. It can be computed as $Q=\sum_{i=1}^{K}(|E_{c_i}|/|E|-(\sum_{x\in c_i} d_x/2|E|)^2)$, where K is the number of communities, c_i denotes community i, $|E_{c_i}|$ is the edges in c_i, and d_x is the degree of node x. The value of Q is in the interval: (-1,1), and a larger value indicates a higher quality. Note that $K=10$ is the default setting in our experiments.

First of all, we illustrate to what extent CEF can improve the community detection performance. To this end, we compared the quality of detected communities before and after using SAM. As can be seen from Fig. 3, if we conduct community detection on the



(a) `Oklahoma` (b) `Enron` (c) `Gowalla`

Figure 3 Performance Improvement Using Community Extraction.



(a) Complete Network (b) Direct Partitioning (c) Using CEF

Figure 4 Community Detection in `Enron` Network

whole networks without using SAM, we obtain the lowest Q values in all three sub-figures. However, as we carefully set τ and σ values in SAM to extract fewer but higher-quality core nodes, the Q values of the detected communities increased gradually, no matter what partitioning method was used. For instance, FN without SAM obtained a Q value lower than 0.4 on the `Gowalla` network; but after using SAM to extract core nodes, the Q value increased rapidly, and finally reached 0.85 on a set of high-quality core nodes. These results clearly demonstrate that by focusing on the core nodes extracted by SAM, CEF indeed improves the performance of community detection greatly.

We then take a closer look at the extracted communities. The `Enron` network is used here as an example. As can be seen from Fig. 4(a), the whole graph of `Enron` consists of a intertwined core and lots of periphery nodes. If we employ CLUTO directly on this network, we obtain ten obscured communities that cannot been identified even with great efforts, as shown in Fig. 4(b). However, if CEF is used instead with $\tau = 0.5$ and $\sigma = 0.02\%$, only 1766 high-quality core nodes will be extracted from the original network, which display a clear structure of ten communities in Fig. 4(c).

4. Conclusions

In this paper, we proposed a novel *structural approximation* concept, upon which a new framework for community extraction was established and validated empirically. We believe community extraction will become a promising solution particularly for the analysis of massive social networks.

Acknowledgments

This research was partially supported by National Natural Science Foundation of China (NSFC) under Grants 71072172 and 61103229. J. Wu was supported in part by NSFC under Grants 70901002 and 71171007.

Note: For space concern, the reference list was moved to the file of supplement information.

References

[1] S. P. Borgatti and M. G. Everett. Models of core/periphery structures. *Social Networks*, 21:375–395, 1999.

[2] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *SIGKDD 2011*, pages 621–628, 2011.

[3] A. Clauset. Finding local community structure in networks. *Physical Review E*, 72:026132, 2005.

[4] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

[5] I.S. Dhillon, S. Mallela, and D.S. Modha. Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 89–98, 2003.

[6] S. Fortunato. Community detection in graphs. *Physics Reports*, 2010.

[7] M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of The Royal Statistical Society Seies A*, 127(2):301–354, 2007.

[8] George Karypis. Cluto — software for clustering high-dimensional datasets, version 2.1.1. Oct. 2007.

[9] George Karypis. METIS and ParMETIS. In *Encyclopedia of Parallel Computing*, pages 1117–1124. 2011.

[10] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[11] M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066113, 2004.

[12] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the USA (PNAS)*, 103(23):8577–8582, 2006.

[13] M.E.J. Newman and M. Girvan. Finding and evaulating community structrue in networks. *Physical Review E*, 69(2):026113, 2004.

[14] L. Tang, X. Wang, and H. Liu. Community detection via heterogeneous interaction analysis. *Data Mining Knowledge Discovery*, 25:1–33, 2012.

[15] Lei Tang and Huan Liu. *Community Detection and Mining in Social Media*. Morgan & Claypool, 2010.

[16] Amanda L. Traud, Eric D. Kelsic, Peter J. Mucha, and Mason A. Porter. Community structure in online collegiate social networks. arXiv:0809.0960, 2008.

[17] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[18] Y. Zhao, E. Levina, and J. Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences of the USA (PNAS)*, 108(18):7371–7326, 2011.