Expert Systems with Applications 42 (2015) 8349–8360

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A novel disjoint community detection algorithm for social networks based on backbone degree and expansion

Yunfeng Xu [a,b], Hua Xu [b,*], Dongwen Zhang [a,b]

[a] College of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China
[b] State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Available online 7 July 2015

Keywords:
Community detection
Social network
Expansion
Conductance

ABSTRACT

Community detection in social networks is a key point to discover the functions and structure of social networks. A great deal of work has been done for overlapping community detection and disjoint community detection, and numerous techniques such as spectral clustering, modularity maximization, random walks, differential equation, and statistical mechanics are used to identify a community in networks, but most of these work adopts pure mathematic and physical methods to discover communities from social networks, on the contrary ignoring the social and biological properties of communities and social networks. In this paper, firstly we propose the community forest model based on these social and biological properties to characterize the structure of real-world large-scale networks, secondly we mainly define a new metric named backbone degree to measure the strength of the edge and the similarity of vertices and give a new sense definition to community based on expansion, thirdly we develop a novel algorithm that based on backbone degree and expansion to discover disjoint communities from real social networks. This algorithm has better performance and effects compared with CNM and GN algorithms in computational cost and visibility. It has worked well on Email-Enron, American College Football, karate club etc. data sets.

1. Introduction

With the advent of massive social networks, community detection in large-scale social networks becomes increasingly important since it may help discover hidden knowledge in large social networks. The hidden knowledge include: structure, function, law, etc. The structure of networks is fundamental interest in large scale network systems, because of their functional implications—commu nities in a social network, for instance, may indicate factions, interest groups, or social divisions; communities in a metabolic network might correspond to functional units, cycles, or circuits that perform certain tasks (Newman, 2013). Community detection is used typically as a tool for discovering and understanding the large-scale structure of networks (Easley & Kleinberg, 2010). Over the past decade community detection (also sometimes called graph partitioning) has been applied to many real-world areas such as biological networks, web graphs, VLSI design, social networks, and task scheduling.

Many algorithms about community detection have been proposed to divide the network into communities, some of them typically choose objective functions that characterize the feature of community, then to optimize them, but these objective functions is typically NP-hard to optimize exactly (Arora, Rao, & Vazirani, 2009; Schaeffer, 2007). Some of them adopt heuristics(Girvan & Newman, 2002; Karypis & Kumar, 1998) or approximation (Leighton & Rao, 1999; Newman, 2013; Spielmat & Teng, 1996) algorithms to optimize some objective functions approximately, these objective functions interpret the community in the real world. But most of these algorithms focus on methodology how to divide the network into communities with Laplacian matrix and eigenvalue resolver, they are elegant but fails to process large-scale networks efficiently and exactly, and they adopt pure mathematic and physical methods such as spectral clustering, modularity maximization, Random walks, differential equation, and statistical mechanics, on the contrary ignoring the social and biological properties of community and network. The social and biological characteristics of community and network refers to the characteristics that got from the study of microscopic feature of the specific network. For example, there are many social properties about social network: weak and strong link, bridge, shortcut,

* Corresponding author.
E-mail addresses: 386839300@qq.com (Y. Xu), xuhua@tsinghua.edu.cn (H. Xu), zdwwtx@163.com (D. Zhang).

http://dx.doi.org/10.1016/j.eswa.2015.06.042
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

neighborhood overlap (Easley & Kleinberg, 2010), authority weight, hub weight (Kleinberg, 1999), K-component and so on. If we consider a social network as a forest, communities in the forest as trees, shrubs, grass, then there are many features are similar between the social networks and the forest. For example, social network is growing, and communities are growing, this feature is like a forest. We can consider these features are the biological characteristics of community and network.

If we mine sociological and biological characteristics about the networks in-depth, we can get a more simple model to characterize the community and the network, then get a efficient algorithm based on this model? Newman discussed the structure of real-world large-scale networks with a look at component sizes, considered the structure of most networks is of a large component filling most of the network, sometimes all of it, and perhaps some other small components that are not connected to the bulk of the network (Newman, 2009). We can consider the real-world large-scale networks is consist of some k-components. The higher the K, K-components have more connectivity, more like a community. A K-component is a k-core, that is like a weak community, but k-component is not a whole community, because all vertices have degree less than k have been removed, then k-component can describe the structure of real-world large-scale networks partly, but not all. So we need a new model to characterize the structure of real-world large-scale networks. We consider many features about social network, then propose the community forest model and an efficient algorithm based on the community forest model.

The contributions of our work are three fold. Firstly we propose the community forest model based on these social and biological properties to characterize the structure of real-world large-scale networks. Secondly we mainly define a new metric named backbone degree to measure the strength of the edge and the similarity of vertices and give a new sense definition to community based on expansion. Thirdly we develop a novel algorithm that based on backbone degree and expansion to discover communities from real social networks.

The rest of paper is organized as follows. Section 2 is an introduction to related work. In Section 3, we formally introduce several concepts related to social networks and the clustering problem. In Section 4, we systematically develop the backbone algorithm. Section 5 is experiment study and Section 6 concludes this study.

2. Related work

In this section we survey related work firstly, then give an emphasis on background and motivation about this paper.

2.1. Related work

A great deal of work has been devoted to detect communities in large-scale networks, they can be categorized into two big classes according to the criterion of whether to allow overlapping: overlapping community detection and disjoint community detection. Overlapping community detection algorithms are reviewed and categorized into five classes: Clique Percolation, Line Graph and Link Partitioning, Local Expansion and Optimization, Fuzzy Detection, Agent-Based and Dynamical Algorithms (Xie, Kelley, & Szymanski, 2013), they are investigated based on the consensus that people in a social network are naturally characterized by multiple community memberships. For the detail of overlapping community detection algorithms, see Xie et al. (2013). Disjoint community detection algorithms are reviewed and categorized into five research lines (Leskovec, Lang, & Mahoney, 2010), they used numerous techniques such as spectral clustering, modularity maximization, random walks, differential equation, and statistical mechanics to identify a community as a set of nodes that has more and/or better links between its members than with the remainder of the network (Leskovec et al., 2010). For the detail of disjoint community detection algorithms, see Leskovec et al. (2010).

Some of the above algorithms give some inspiration to our research work. Very relevant to our work is that of Newman (2004) and Clauset, Newman, and Moore (2004), Kleinberg (1999), Leskovec et al. (2010), Kannan, Vempala, and Vetta (2004) and Palla, Derényi, Farkas, and Vicsek (2005).

Newman et al. analyze a hierarchical agglomeration algorithm and describe a community concept depending on the modularity of the communities. Kleinberg propose an algorithmic formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of hub pages that join them together in the link structure (Kleinberg, 1999). We integrated these two concepts of authority and hub as one concept named network weight in undirected network. Leskovec et al. (2010) defined the network community profile (NCP) that characterizes the quality of network communities as a function of their size. Inspired by the network community profile (NCP), we use a scatter diagram to describe the quality of the whole community discovery algorithm, a scatter diagram of the *X* axis is community size, the vertical axis is the conductance and expansion.

Kannan et al. denoted the expansion formally. The expansion of a community is the minimum ratio over all cuts of the community of the total (Kannan et al., 2004). In this paper, we find expansion should gradually decreases from the center of the community to the boundary of the community, we use this feature and backbone degree to add new nodes to the community gradually from the center of the community, until the expansion of the community began to grow bigger, this process can divide communities from social networks.

Kannan et al. proposed a metric named conductance that provides the measure of the quality of an individual cluster (Kannan et al., 2004). Conductance is a expansion-like property, Leskovec et al. consider conductance to be a good metric that characterizes the quality of network communities (Leskovec et al., 2010). In this paper we use the conductance and the expansion as the main evaluation metric to the community detection algorithm.

Easley and Kleinberg (2010) defined closure, structural holes, weak and strong link, bridge, shortcut, neighborhood overlap, etc. The neighborhood overlap characterizes the strength of an edge (Easley & Kleinberg, 2010). On the basis of these concepts, we put forward the metric backbone degree to characterizes the strength of the links between nodes and communities.

Palla et al. (2005) proposed the clique percolation method, their community definition relies is that a typical community consists several complete(fully connected) subgraphs that tend to share many of their nodes, more precisely, a k-clique-community as a union of all k-cliques (complete subgraphs of size *k*) that can be reached from each other through a series of adjacent k-cliques (where adjacency means sharing $k-1$ nodes). This method is based on first locating all cliques (maximal complete subgraphs) of the network and then identifying the communities by carrying out a standard component analysis of the clique-clique overlap matrix. In this paper we propose the notion of neighborhood overlap to measure the strength between tow vertices, this notion coincides with the count of 3-clique between two vertices.

We studied most of above algorithms about overlapping community detection and disjoint community detection. Although the area of overlapping community detection is the hot area currently, but we consider that there is still a big space to improve and broad application prospects in the area of disjoint community detection. Our current work mainly focuses in the area of disjoint community detection in undirected networks, we will explore the area of overlapping community detection lately.

2.2. Research motivation

Definition of the notion of community decides how to found community in the network, then how to define the notion of community? The intuition of community is a set of vertices that connections between the vertices are denser than connections with the rest of the network (Leskovec et al., 2010; Radicchi, Castellano, Cecconi, Loreto, & Parisi, 2004). Radicchi proposed the notion of community quantitatively: in a strong community each vertex has more connections within the community than with the rest of the graph; in a weak community the sum of all degrees within the community is larger than the sum of all degrees toward the rest of the network (Radicchi et al., 2004). Luccio and Sami proposed the notion of community called minimal groups in 1969, Lawler renamed them LS sets in 1973 (Radicchi et al., 2004). LS set is like strong community. Another definition is called k-core, a k-core of a graph G is a maximal connected sub graph of G in which all vertices have degree at least k (Seidman, 1983). K-core is like weak community. From discussed above, we found that those notions of community are concise and clear, but not to be visualized, and no detailed depiction to internal structure. Is there a notion can visualize the notion of community? Which means that the notion can give clear boundary and internal structure.

Along with the development of social network research in recent years, people put forward a lot of new concepts about social network structures such as weak and strong link, bridge, shortcut, neighborhood overlap, etc. If we analysis the connection between the vertices in a network, we find those links include: weak and the strong link, bridge, shortcut and so on. If the contact as branches of the trees, so a community can be seen as a tree, a social network can be as a community forest, the network contains strong communities is a community forest, other contains weak communities is bushes. If according to this hypothesis, we can give more biological properties to the community, and redefine the notion of community. Integrated the concept of the social network, this paper proposed a new metrics—backbone degree, and redefined the concept of community, and proposed the community forest model to visualize the concept of community.

3. Model and problem formalization

In this section, we propose the community forest model and define the problem of community detection and introduce several related concepts and necessary notations.

3.1. Community forest model

Social network and forest have some similar features in morphology and structure. Fig. 1 is a visualization of a social network with more than 30,000 users, its like a dense forest. There are some boundary communities around the giant component. Communities in social networks always consist of core vertices, core backbones and boundary vertices, their morphology and structure are similar with trees, shrubs and grass in forest. Fig. 2 is a visualization of six subgraphs, some are sparse that like shrubs and grass, some are dense that like trees. If we can consider a social network as a forest, communities in the forest are trees, shrubs, grass.

Communities in social network have some relations or have no relations, this feature is like these trees, shrubs and grass in the forest. Big communities in social networks can derive new small communities, this feature is like these trees, shrubs and grass in the forest. There are many features are similar between the social networks and the forest. A community is defined as a subset of vertices within the graph such that connections between the vertices are denser than connections with the rest of the network (Radicchi et al., 2004). Why we consider a community like a tree? A tree consists of roots, trunks and leaves, a community consists of vertices and relations. Some relations are strong and some relations are weak; some vertices are core vertices and some vertices are on the border, like leaves. We consider the strong relations are tree trunks, some core vertices are tree roots, some vertices on the border are leaves.

In its most basic form, the problem of community detection in networks is one of dividing the vertices of a given network into nonoverlapping groups such that connections within groups are relatively dense while those between groups are sparse (Newman, 2013). In this paper, discovering communities from

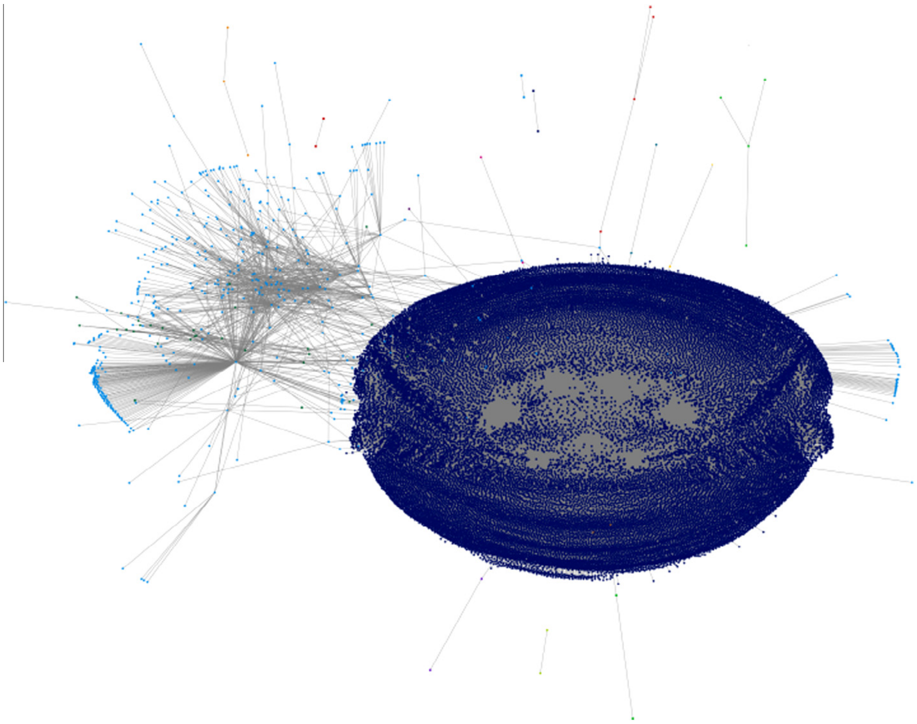

Fig. 1. The visualization of an complete online social network that consisted of more than 30,000 users.

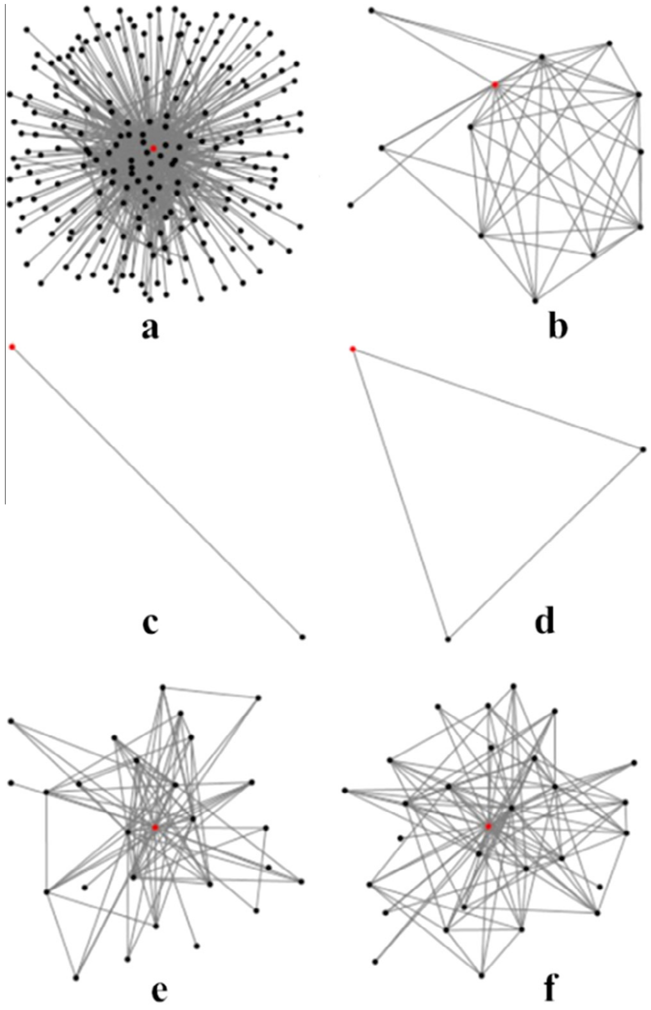

Fig. 2. The visualization of six subgraphs that belong to the network in Fig. 1.

networks is like finding trees from the forest. But how to find a tree from a forest? We only have the edges and vertices, we need a metric to measure the edges, we call the metric as backbone degree. Let a community as a tree, then the edges are the trunks of the tree. A edge consists of two vertices and a relation, so the backbone degree must measure these three factors. The edges like the sections of the bamboo. Every section consists of two joints and a bar. The relations is the bars, The neighborhood overlap measure can be used to represent the strength of the bar. The network weight measure can be used to represent the strength of the joint. If this metric is worked, detecting a community from a network is like this process: Firstly finding the edge with biggest backbone degree, secondly finding nearest vertex based backbone degree until the boundary of this community, repeat the above step in the rest of the vertices, until all the vertices are divided. Control the backbone choice, can make the algorithm more extensibility in operation. For example, if the algorithm allows no backbone is selected in the divided vertices, the community detection is none overlapping, otherwise is overlapping. In this paper, we mainly discuss the non-overlapping community detection.

3.2. Problem formalization

Given an undirected graph $G(V,E)$ with $\mid V \mid$ vertexes and $\mid E \mid$ edges. Let $n =\mid V \mid,\ m =\mid E \mid$. Let C be a set of vertices in a community, where C_n is the number of vertices in $C, C_n =\mid C \mid$. Let $E_C = \{(u,v) \in E : u \in C, v \in C\}$, C_m is the number of edges in $C, C_m =\mid E_C \mid$. Let $C_{BE} = \{(u,v) \in E : u \in C,\ v \notin C\}$, $\mid C_{BE} \mid$ is the number of edges on the boundary of C. Let d_u be the degree of vertex u. Let NB_u be the neighborhood vertices set of vertex u. Let NB_C be the neighborhood vertices set of community $C, NB_C = \{v : (u,v) \in E,\ u \in C,\ v \notin C\}$.

Definition 1 (*Network Weight*). Let the identifier of vertex v be i, the network weight of any vertex in graph G can be represented as x_j. we can use NW_v represent the network weight of v.

$$NW_v = \sum_{j=1}^{n} A_{ij} \frac{x_i}{d_j}$$

The network weights according to the definition of the HIT algorithm (Kleinberg, 1999), but the vertex weights of HITS algorithm needs a lot of calculation to balance, in order to save computation time, a relative weighting of vertices can be considered $x_j = \frac{d_j}{2m}$ then

$$NW_v = \frac{1}{2m} \sum_{j=1}^{n} A_{ij}$$

Definition 2 (*Community Expansion Degree*). This metric measures the number of edges that point outside the community (Kannan et al., 2004).

$$EX_C = \frac{\mid CB_E \mid}{C_n}$$

Definition 3 (*The Difference of Expansion Degree*). The change of expansion degree after joining a new vertex i to community C.

$$DE(i) = EX_{C \bigcap \{i\}} - EX_C$$

Definition 4 (*The probability that the vertex i belongs to the community C*).

$$P(i \in C) = \frac{\mid (NB_i \bigcap C) \mid}{d_i}$$

Definition 5 (*Neighborhood overlap*). Given two vertices u and v, let NB_u be the set of vertices that are the neighborhood of vertex u, let NB_v be the set of vertices that are the neighborhood of vertex v. Let NO_{uv} be the neighborhood overlap of u and v.

$$NO_{uv} = \begin{cases} \frac{|NB_v \bigcap NB_u|}{|NB_v \bigcup NB_u| - 2}, & \text{there is an edge between } u \text{ and } v \\ 0, & \text{there is not an edge between } u \text{ and } v \end{cases}$$

Definition 6 (*Backbone*). A backbone consists of an edge and two vertices that connecting to the edge. If a backbone is connected to the outside of the current community, we call the vertex of the backbone in the current community named interior vertex, another named external vertex.

Definition 7 (*Backbone degree*). The backbone degree of an edge with vertex u and vertex v is:

$$D_{uv} = (NW_u + NW_v) \times NO_{uv} + \delta$$

D_{uv} can measure the strength of the edge and the similarity of nodes. When vertex u and v have no neighborhood, then $NO_{uv} = 0$

,$D_{uv} = \delta$. δ is a constant parameter for smoothing, we let $\delta = 0.01$ based on experience.

Definition 8 (*Max backbone degree of community C*). Let CD_{max} be the maximal backbone degree in community *C*, the backbone with CD_{max} is the core backbone of community *C*.

$$CD_{max} = max\{D_{uv}, u \in C, v \in C\}$$

Definition 9 (*Community in new sense*). This paper gives a new definition of community in a new sense. Community is a set with some vertices that expand outward from the core backbone according the D_{uv} gradually, and the expansion diminishes gradually, until *EX* is minimum.

$$C = \{u : u \in C, v \notin C, (u, v) \in E, EX_{\{C \bigcup v\}} > EX_C\}$$

Definition 10 (*Community forest*). The communities set in *G* can be defined as community forest.

$$CF = \{C : C \in V, v \notin C, EX_{\{C \bigcup v\}} > EX_C\}$$

Definition 11 (*Triadic Closure*). If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future (Rapoport, 1953).

Definition 12 (*Co-friend count*). In a community network, if there is a mutual friend between two people, then they become good friends will likely increase, so take a co-friend count between two vertices to measure the similarity of the two vertices.

Definition 13 (*Member closure*). In a social network, if a person's friends took part in a community, then the possibility to participate in the same community will improve (Rapoport, 1953).

Definition 14 (*Member closure count of vertex i to community C*). If the vertex *i* have *x* friends in the community *C*, the member closure count of vertex *i* to community *C* equals *x*. $x = \mid NB_{\{i\}} \cap C \mid$, the member closures count of vertex *i* can measure similarity to community *C*.

Definition 15 (*The boundary set of community C*). Let C_{BV} be the boundary set of community $C, \mid C_{BV} \mid$ is the number of vertices on the boundary of *C*.

$$BV_C = \{v : (u, v) \in E, u \in C, v \notin C, EX_{\{C \cup v\}} > EX_C\}$$

Definition 16 (*Sum of backbone degree to the edges of vertex v point to community C*). This metric can measure the distance between vertex v and community *C*.

$$NC_v = \sum_{u \in C, v \notin C} D_{uv}$$

Definition 17 (*Nearest vertex to community C*). Let MAX_{NC} is the maximal value in $\{NC_v : v \in \{NB_C - BV_C\}, EX_{\{C \cup v\}} < EX_C\}$, if $NC_v = MAX_{NC}$, vertex v is the nearest vertex to community C.

$$v = \{v : (u, v) \in E, u \in C, v \notin C, v \in \{NB_C - BV_C\}$$

$$EX_{\{C \cup v\}} < EX_C, NC_v = MAX_{NC}\}$$

Definition 18 (*The prediction of relation between vertex i and j for non-overlapping community detection*). NO_{uv} is the backbone degree of pair (i,j), θ_u is a vector that characterized the membership of vertice *u* (Airoldi, Blei, Fienberg, & Xing, 2009), θ_v is a vector that characterized the membership of vertex v, K is the count of all communities in *G*.

$$p(y_{ij} = 1 | \theta_u, \theta_v) = \sum_{v \in NB_i, u \in NB_j} \sum_{k=1}^{k} \theta_{uk} \theta_{vk} NO_{uv}$$

NO_{uv} must be relaxed: when there is not a edge between u and $v, NO_{uv} = 0$, here let $NO_{uv} = \frac{|NB_v \bigcap NB_u|}{|NB_v \bigcup NB_u| - 2}$.

4. Backbone degree algorithm

According to the community forest model, the process of community detection can be defined like that: finding the core backbone of each community and looking for the boundary of each community. If we find the core backbone of each community, then the number of communities in network is determined. After determining the number of communities and the core backbone of each community, our algorithm can be extended to process large-scale networks with parallel computing. Because the space is limited, we will discuss the problem in the next article.

Why beginning from the core backbone, rather than starting from the core vertex? If we consider the community as a tree in the forest, based on the assumption that the community must expand from the core backbone. The backbone with CD_{max} is the core backbone of community *C*. when community *C* contains only the core backbone edge,

$$EX_C = \frac{\mid C_{BE} \mid}{C_n} = \frac{d_u + d_v - 2}{2} u \in C, \quad v \in C$$

then $d_u + d_v = 2(EX_C + 1)$ and

$$\begin{aligned} D_{uv} &= (NW_u + NW_v) \times NO_{uv} + \delta \\ &= \frac{d_u + d_v}{2m} \times NO_{uv} + \delta \\ &= \frac{EX_C + 1}{m} \times NO_{uv} + \delta \end{aligned}$$

where *m* and δ are constant, EX_C and NO_{uv} are variable, $m = \mid E \mid$, we let $\delta = 0.01$ for smoothing based on experience, D_{uv} integrates EX_C and NO_{uv}, in order to more accurately choose the core of the community, and this avoids those nodes such as structural holes. The greater NO_{uv}, the denser the community internal connection, this point coincides with the one proposed in Radicchi et al. (2004) in the framework of the identification of communities. If we discover community since core vertex, the only available metric to determine the core vertex is the weight of vertex, the weight of core vertex in undirected network is closely related to the degree of core vertex, but in social network, a vertex with big degree also may be a structure hole. If we let a backbone with the largest backbone degree as the core of the community, that will avoid the structure hole, because the backbone degree insists of the weight of the 2 vertices and the neighborhood overlap, so if a backbone has the great backbone degree, the backbone is likely the core of a community or near the core of a community.

Our algorithm firstly calculates the backbone degree of each backbone in the social network, and saves these backbone degree to a backbone list, then sorts the backbone list in descending order. Let initial community be empty, selecting the backbone with the largest backbone degree in the list as the initial backbone to the current community, then adding the backbone with the largest backbone degree in the set that connected to the current community in turn. If the expansion is smaller after adding a new backbone to the current community, then continually adding the backbone with the largest backbone degree in the set that connected to the current community, else add the external vertex of the backbone to the boundary set of the current community, continue to find the vertex with the maximum backbone degree that connected with the current community until there is no longer eligible vertex in the neighbor set of the current community, a new community is divided completely right now. In accordance with the above method to continue the iteration, divided the rest vertices into new communities, until there are no backbones that their backbone degree is greater than the threshold value *f* in backbone list, or the count of the rest vertices less than parameter *w*. Here *w* is according to $|V|$, for example, $w = \frac{|V|}{10}$, because in large-scale social network, when the rest vertices in the social network are less, there are not more truly valuable communities in the rest vertices, if to use the above steps again, will find out the very small and useless community, at this time the rest of the vertices can be collected with some simpler algorithms, such as using member closure to determine that a vertex is belong to which community. Backbone degree threshold *f* can be fixed by experience or requirements of user, such as users want to find these communities with the core backbone that backbone degree is more than 0.3, then let $f = 0.3$.

4.1. Framework of backbone degree algorithm

Given an undirected graph $G(V,E)$ with $|V|$ vertices and $|E|$ edges, given the node list *NL* to save the vertices in *V*, let the current community is C_i, the neighbor set of C_i is NB_{C_i}, the boundary set of C_i is BV_{C_i}, given the backbone list *BL* to save the backbones in *E*. Backbone degree algorithm implementation is shown in Algorithm 1.

Algorithm 1. Backbone degree algorithm implementation

```
Data: An undirected $G(V,E)$. Result: The community set CF in
  G.
begin
1    $NL \Leftarrow V, BL \Leftarrow$ edges with backbone degree $\geqslant f$ in $E, CF \Leftarrow null$,
     $i \Leftarrow 0$. Sort $BL$ according to descending order, $index_{BL} \Leftarrow 0$.
2    Get a backbone $b$ from $BL$ according to $index_{BL}, index_{BL}++$,
     get vertices $u$ and $v$ from $b$, note the backbone degree of $b$
     as $BD_b$,
     note the size of $NL$ as $nl$.
3    while $BD_b \geqslant f$ and $nl \geqslant w$ do
4      if $u \in NL$ and $v \in NL$ then
5        $C_i \Leftarrow \{u, v\}$
6        $E_C\_PRE \Leftarrow$ the Expansion degree of $C_i$ .
7        calculate the $NB_{C_i}$ of $C_i, BV_{C_i} \Leftarrow null$.
8        if $\{NB_C - BV_C\} = Null$, then
9          add $C_i$ to $CF; i++$; goto step2.
10       else
11         find the nearest vertex $nv$ from $\{NB_{C_i} - BV_{C_i}\}$
           based on
         backbone degree, add vertice $nv$ to $C_i$, calculate the
         Expansion
         degree of $C_i$ and note it as $E_C\_cur$.
12         if $(E_C\_cur - E_C\_PRE < 0)$, then
13           remove vertice $nv$ from $NL$ and add vertice $nv$ to $C_i$,
             goto step11.
14         else delete vertice $nv$ from $C_i$, add vertice $nv$ to $BV_C$,
15           if $\{NB_C - BV_C\} = Null$, then
16           add $C_i$ to $CF, i++$, goto step2.
17           else
18             goto step11.
19           end if
20         end if
21       end if
22     else
23       goto step2;
24     end if
25   end while
26     Collect all vertices that divided into no community or
     several communities.
27     return $CF$.
end
```

4.2. Algorithm time complexity

Our algorithm uses merge sort to sort the backbone list, it runs in time $O(m \log m)$, and the process of discovering community runs in time $O(n + m)$, so our algorithm runs in time $O(m \log m + n + m)$ for a network with n vertices and m edges. Because not all of the backbones are the core backbones, if we filter the backbone list according to a threshold *f*, the count of the backbone list will fall sharply, $O(m \log m)$ will fall sharply too, so our algorithm runs in time $O(n + m)$ approximately. We analyzed backbone degree of five data set. Table 2 is the edges with biggest backbone degree. Table 3 is the count of edges with $f \geqslant 0.2$ and $f \geqslant 0.3$. We found that the biggest backbone degree is 1.282727, the minimum backbone degree is 0.01, when *f* values change, the count of the backbone list will change sharply, that is shown in Table 3.

4.3. Algorithm comparison

CNM algorithm runs in time $O(md \log n)$ for a network with *n* vertices and *m* edges where *d* is the depth of the dendrogram. Girvan Newman algorithm runs in time $O(n3)$. Backbone degree algorithm runs in time $O(n + m)$ approximately, but Girvan Newman algorithm and CNM algorithm attempt at optimizing the modularity that is different to backbone degree algorithm. So it is hard to say which algorithm is better in time complexity.

The goal of backbone degree algorithm is detecting communities from networks based on backbone degree and community forest model, and it is scalable and can be applied to large-scale social networks. Backbone degree algorithm can discover networks in different depth based on backbone degree threshold *f* and parameter *w*, this made it very flexible, and it can predict the relation between every vertices *i* and *j* in networks for disjoint and overlapping community detection.

5. Experiments

In this section, we study the effectiveness and accuracy of backbone degree algorithm and compare it with CNM (Clauset et al., 2004) algorithms mainly, also we compare our algorithm with Martin Rosvall's algorithm (Rosvall & Bergstrom, 2008) and GN algorithm simply. CNM algorithm implementation is from Stanford Network Analysis Platform (SNAP). SNAP is a general purpose network analysis and graph mining library. It is written in C++ and easily scales to massive networks with hundreds of millions of nodes, and billions of edges (Leskovec, 2014).

5.1. Data set

We use a artificial network and some standard data set: Zacharys Karate Club, American College Football, Email-Enron data

Table 1
Data set description.

Data set	Vertices	Edges	Known communities
An artificial network	19	21	3
Zachary's Karate Club	34	78	2
American College Football	115	613	12
Enron email communication network	36,692	183,831	Unknown
DBLP computer science bibliography	317,080	1,049,866	13,477

Table 2
The edges with biggest backbone degree.

Data set	Biggest backbone edge	Backbone degree
An artificial network	Anyone	0.01
American College Football	763,689	1.282727
Zachary's Karate Club	3331	1.01346
Enron email communication network	76,136	0.392861
DBLP computer science bibliography	55885,286328	0.660222

Table 3
The count of edges with $f \geqslant 0.2$ and $f \geqslant 0.3$.

Data set	$f \geqslant 0.2$	$f \geqslant 0.3$
An artificial network	None	None
American College Football	449	411
Zachary's Karate Club	31	9
Enron email communication network	106,592	12
DBLP computer science bibliography	625,721	12,536

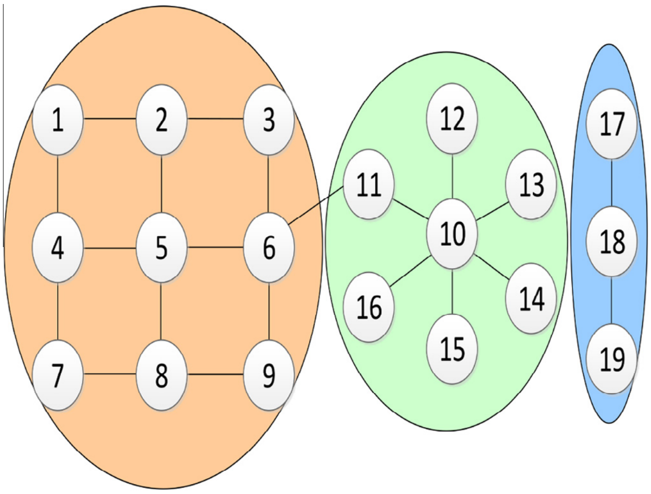

Fig. 3. An artificial network that consisted of star, mesh and line topology structure.

Table 4
The backbone degree of edges in the artificial network.

Edge name	Backbone degree	Belonged to structure
1,2	0.01	Mesh
1,4	0.01	Mesh
2,3	0.01	Mesh
2,5	0.01	Mesh
3,6	0.01	Mesh
4,5	0.01	Mesh
4,7	0.01	Mesh
5,6	0.01	Mesh
5,8	0.01	Mesh
6,9	0.01	Mesh
6,11	0.01	None
7,8	0.01	Mesh
8,9	0.01	Mesh
10,11	0.01	Star
10,12	0.01	Star
10,13	0.01	Star
10,14	0.01	Star
10,15	0.01	Star
10,16	0.01	Star
17,18	0.01	Line
18,19	0.01	Line

Table 5
The result of dividing the artificial network with our algorithm.

Vertex number	Belonged to structure	Neighbor vertices	Community ID
1	Mesh	2, 4	2
2	Mesh	1, 3, 5	2
3	Mesh	2, 6	2
4	Mesh	1, 5, 7	2
5	Mesh	2, 4, 6, 8	2
6	Mesh	3, 5, 9, 11	2
7	Mesh	4, 8	2
8	Mesh	5, 7, 9	2
9	Mesh	6, 8	2
10	Star	11, 12, 13, 14, 15, 16	1
11	Star	6, 10	1
12	Star	10	1
13	Star	10	1
14	Star	10	1
15	Star	10	1
16	Star	10	1
17	Line	18	0
18	Line	17, 19	0
19	Line	18	0

set, DBLP collaboration network. American College Football and Karate Club are standard data sets to prove the validity of community detection algorithm, DBLP collaboration network is a ground-truth network, the detailed description of these data sets is shown in Tables 1–3.

The artificial network is a sparse network that consisted of star, mesh, line topology, and the backbone degree of very edge in the artificial network is 0.01, the network is very spare and specific, that is shown in Fig. 3 and Table 4, it is created by us for demonstrating our algorithm.

Zachary's Karate Club is a social network of friendships between 34 members of a karate club at a US university in the 1970s. Wayne Zachary observed social interactions between the members of a karate club at an American university. He built network of connections with 34 vertices and 78 edges in the early 1970s. By a chance, a dispute arose between the club's administrator and the karate teacher, the club split into two small communities with the administrator and the teacher being as the central persons.

American College Football is a network of American football games between Division IA colleges during regular season Fall 2000.

Enron email communication network covers all the email communication within a data set of around half million emails (Leskovec, Lang, Dasgupta, & Mahoney, 2009). This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. Nodes of the network are email addresses and if an address *i* sent at least

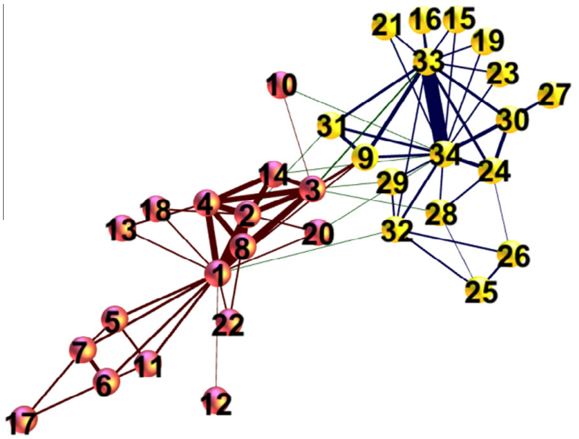

Fig. 4. The result of applying backbone degree algorithm to Zachary Karate Club network, the yellow is community 0, the red is community 1. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

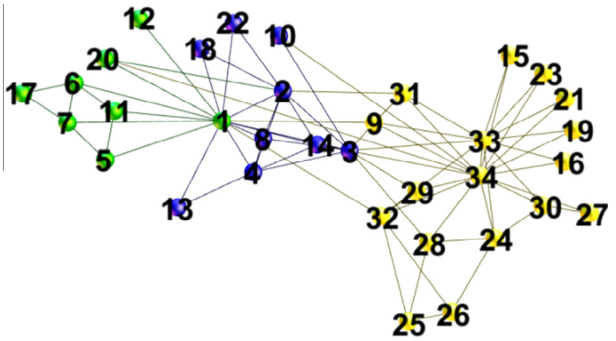

Fig. 5. The result of applying CNM to Zachary Karate Club network. The yellow is community 0, the blue is community 1, the green is community 2. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)

one email to address j, the graph contains an undirected edge from i to j.

The DBLP computer science bibliography provides a comprehensive list of research papers in computer science. In this paper we construct a co-authorship network where two authors are connected if they publish at least one paper together (Yang & Leskovec, 2012).

5.2. An artificial network test

We apply our algorithm to an specific artificial network that consisted of star, mesh, and line topology that shown in Fig. 3. This network is a very spare network, and the backbone degree of edges in it is 0.01 when $\delta = 0.01$. Table 4 are the backbone degree of Edges in the Artificial network. Table 5 is the result of dividing the artificial network with backbone degree algorithm. The result of CNM algorithm is same to Table 5. This test shows that our algorithm still has the very good adaptability for some sparse networks with very small backbone degree.

5.3. Zachary Karate Club test

We apply our algorithm to karate club networks, Fig. 4 is the result of applying our algorithm to Zachary Karate Club network. Our algorithm divides this network in two communities clearly. Expansion and conductance of applying our algorithm to Zachary Karate Club network are shown in Table 7. Fig. 5 is the result of applying CNM to Zachary Karate Club network, CNM divides this network in three communities, there is no clear boundary and internal structure between communities 1 and 2. Expansion and conductance of applying CNM algorithm to Zachary Karate Club network are shown in Table 6. Results of Table 7 are significantly better than those in Table 6. GN (Girvan & Newman, 2002) algorithm divides this network in 5 communities. Martin Rosvall's algorithm (Rosvall & Bergstrom, 2008) divides this network in 6 communities. The results of GN algorithm and Martin Rosvall's algorithm make some sense to their models, but they are not same with the standard result completely.

Table 6
The result of dividing Zachary Karate Club with CNM algorithm.

Community ID	Community size	Conductance	Expansion
1	9	0.380952	1.777778
2	8	0.333333	1.5
0	17	0.128205	0.588235

Table 7
The result of dividing Zachary Karate Club with backbone degree algorithm.

Community ID	Community size	Conductance	Expansion
0	17	0.128205	0.588235
1	17	0.128205	0.588235

Table 8
Backbone degree algorithm implementation process to Zachary Karate Club.

Vertex ID	Current expansion	Community ID	Joining order	MAX_{NC}
34	13.5	0	1	1.013
33	13.5	0	1	1.013
9	9.333	0	2	0.414
31	6.5	0	3	0.544
30	5.2	0	4	0.408
24	4.166	0	5	0.624
32	3.857	0	6	0.233
27	3.125	0	7	0.207
29	2.666	0	8	0.177
28	2.4	0	9	0.177
19	2	0	10	0.165
23	1.666	0	11	0.165
21	1.384	0	12	0.165
15	1.143	0	13	0.165
16	0.933	0	14	0.165
25	0.8125	0	15	0.108
26	0.588	0	16	0.236
10	0.555	0	17	0.01
2	11.5	1	1	0.653
1	11.5	1	1	0.653
4	8.333	1	2	0.844
3	7.25	1	3	1.153
8	5	1	4	1.189
14	3.666	1	5	1.134
9	3.286	1	6	0.249
31	2.875	1	7	0.232
13	2.333	1	8	0.185
22	1.9	1	9	0.171
18	1.545	1	10	0.171
20	1.333	1	11	0.168
5	1.307	1	12	0.159
11	1.143	1	13	0.287
7	1.067	1	14	0.27
6	0.875	1	15	0.515
17	0.706	1	16	0.255
12	0.579	1	17	0.01

We display the our algorithm implementation process to Zachary Karate Club in Table 8, this process does not include step 26. The process of dividing karate club is shown in Table 8, we found that community 0 and 1 are overlapping on vertex 9, 31, 10 before

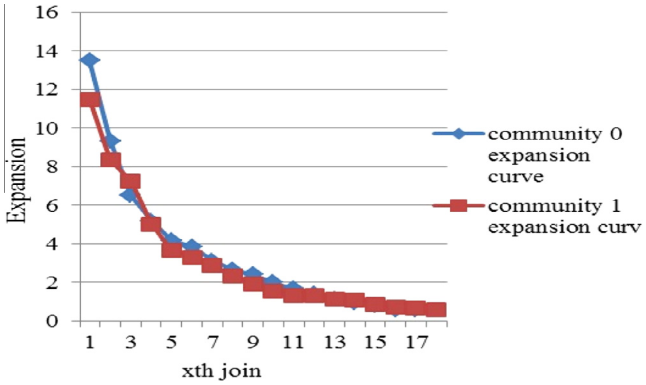

Fig. 6. The curve of expansion.

step 26. If we run step 26, the result is shown in Fig. 4, and it is same with the standard result completely. We also found that the expansion of community decreases gradually with the vertex joining, the curve is shown in Fig. 6, and the vertices's order of joining community is based on backbone degree. This phenomenon is fully verified the definition of community in this paper. Tracking the vertices's order and their backbone degree to see Fig. 4. The MAX_{NC} of Table 8 is defined in Definition 17.

5.4. American College Football

American College Football is a network of American football games between Division IA colleges during regular season Fall 2000.

Backbone degree algorithm can divide the network into 12 communities exactly, CNM algorithm divides the network into 5 communities. The result of applying backbone degree algorithm to American College Football network is shown in Fig. 7. In this result, six communities compare with the standard data set completely consistent, one community is less one vertex than standard data set, two communities are more 2 vertices than standard data set, one community is different with standard data set in 3 vertices, two communities are different with standard data set in 5 vertices. The result of applying CNM algorithm to American College Football network is shown in Fig. 8. The comparison of expansion and conductance is shown in Tables 9 and 10. From discussed above, the result of backbone degree algorithm is better than the result of CNM algorithm, because although the conductance and expansion score of CNM algorithm are good, but the result of CNM algorithm can not reflect the real structure of the American College Football network. This proved backbone degree algorithm can discover the structure of social networks exactly.

5.5. Email-Enron data set

Enron email communication network covers all the email communication within a data set of around half million emails (Leskovec et al., 2009). This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. Nodes of the network are email addresses and if an address i sent at least one email to address j, the graph contains an undirected edge from i to j.

The conductance scatter diagram of applying CNM algorithm and our algorithm to Email-Enron network is shown in Fig. 9. In this result, we find our algorithm is slightly better than CNM

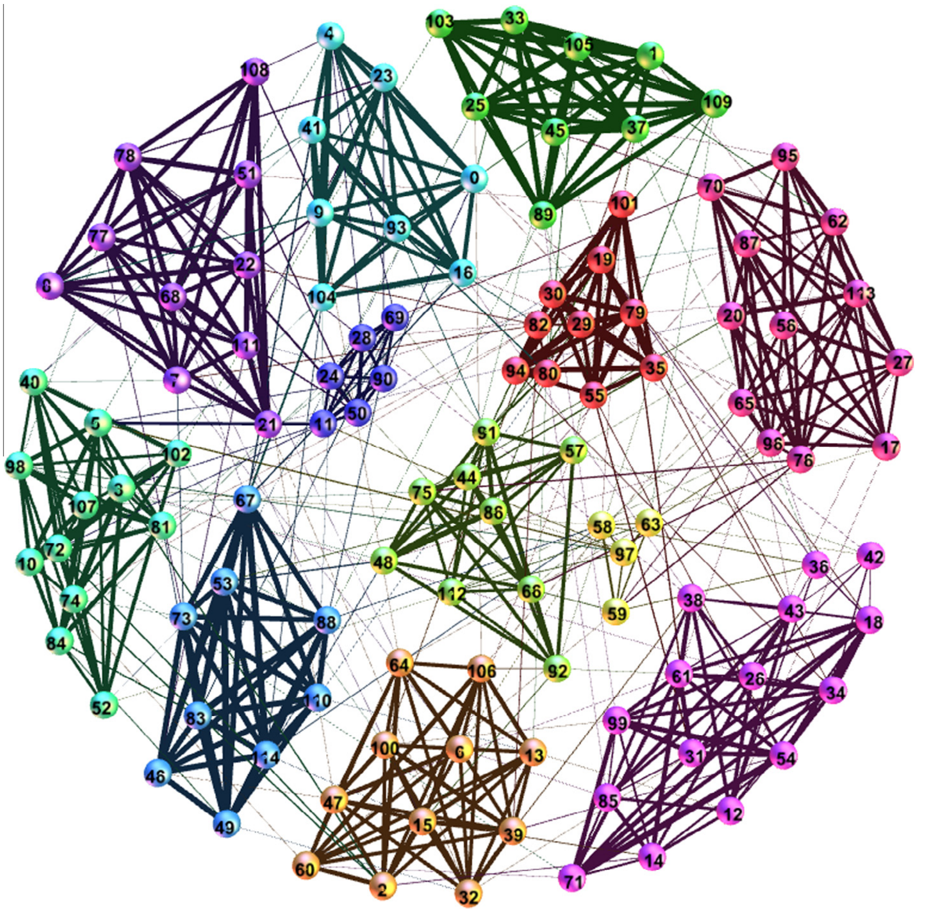

Fig. 7. The result of applying backbone degree algorithm to American College Football network.

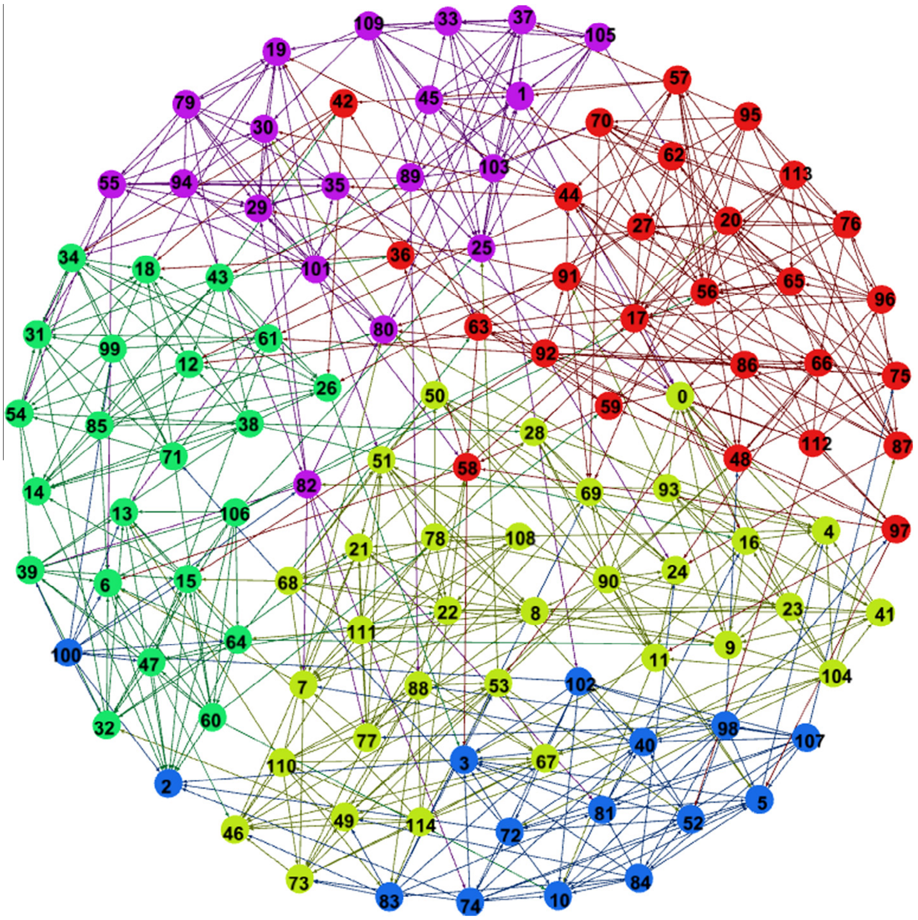

Fig. 8. The result of applying CNM algorithm to American College Football network.

Table 9
The result of dividing American College Football with backbone degree algorithm.

Community ID	Community size	Conductance	Expansion
0	9	0.258	2.778
1	10	0.352	3.800
2	9	0.294	3.333
3	10	0.273	3.000
4	15	0.240	2.400
5	8	0.364	4.000
6	9	0.354	3.778
7	12	0.262	2.833
8	12	0.250	2.667
9	11	0.290	3.273
10	6	0.483	4.667
11	4	0.657	5.750

Table 10
The result of dividing American College Football with CNM algorithm.

Community number	Community size	Conductance	Expansion
0	19	0.239024	2.578947
1	32	0.193084	2.09375
2	15	0.329268	3.60000
3	22	0.262712	2.818182
4	27	0.218978	2.222222

algorithm in conductance. The conductance of our algorithm is mostly low than CNM algorithm, and it is more compact and stable.

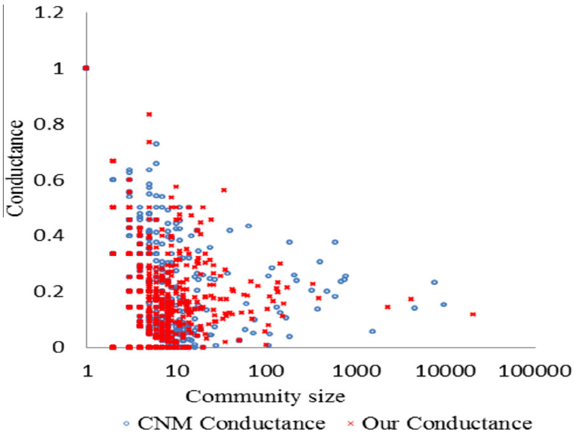

Fig. 9. The conductance scatter diagram of applying CNM algorithm and our algorithm to Email-Enron network.

The expansion scatter diagram of applying CNM algorithm and our algorithm to Email-Enron network is shown in Fig. 10. In this result, we find our algorithm is slightly less than CNM algorithm in conductance. The expansion of our algorithm is mostly high than CNM algorithm. But it is more compact and stable than CNM algorithm.

We find that backbone degree algorithm and CNM algorithm have the same distribution generally in conductance and expansion. On the other hand these result proved that the structure of Email-Enron network that found by backbone degree algorithm

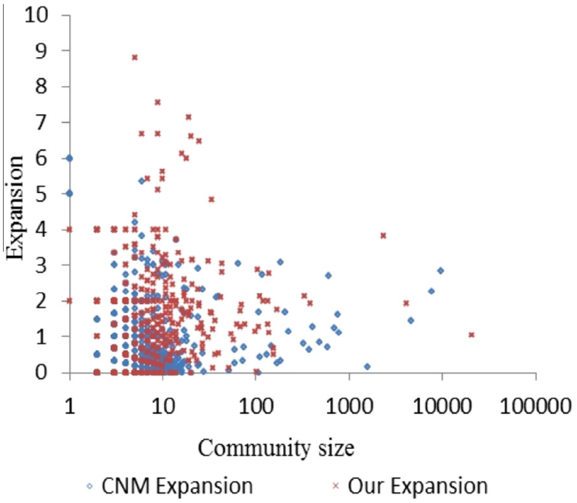

Fig. 10. The expansion scatter diagram of applying CNM algorithm and our algorithm to Email-Enron network.

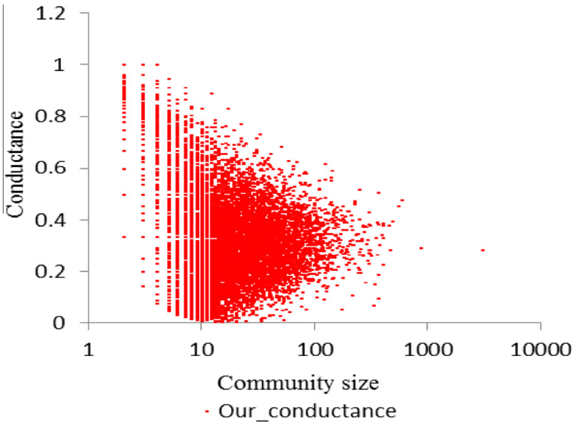

Fig. 11. The conductance scatter diagram of applying backbone degree algorithm to DBLP network.

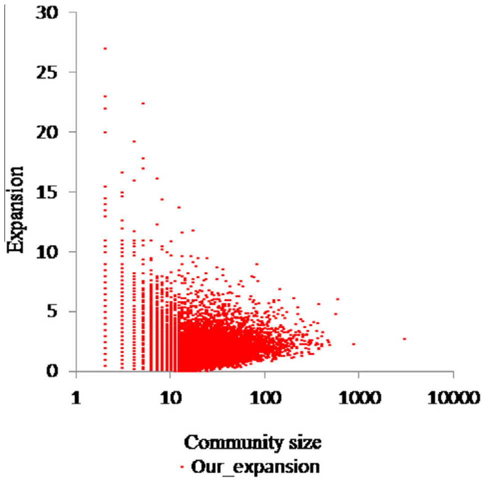

Fig. 12. The expansion scatter diagram of applying backbone degree algorithm to DBLP network.

is very close to CNM algorithm found. In the large-scale social networks it is hard to say which algorithm is better, because in addition to the relationship between social network data, there are many other implicit factors.

5.6. DBLP collaboration network

The conductance scatter diagram of applying backbone degree algorithm to DBLP network is shown in Fig. 11. The horizontal axis is the community size, the vertical axis is the conductance of community. As community size increases, the conductance decreases gradually and tends to be stable at about 0.3. And most of communities size between 10 and 100 in DBLP networks, this result conforms to a common sense that most of the social circle under 100 people.

The expansion scatter diagram of applying backbone degree algorithm to DBLP network is shown in Fig. 12. The horizontal axis is the community size, the vertical axis is the expansion of community. As community size increases, the expansion decreases gradually and tends to be stable at about 2. We find those communities under 10 vertices that with greater expansion, this is same as CNM algorithm in Email-Enron networks. It is a little defect of backbone degree algorithm, we will make up for the defects in the later work.

Because CNM algorithm requires more memory than backbone degree algorithm, we cannot get the result of CNM in DBLP collaboration network based on our current computing environment.

5.7. Discussion

Through above experiments, we found that CNM algorithm can get good score of conductance and expansion, but not accurate to find the structure of social networks on small data sets. So conductance and expansion are reference to measure the result of community detection, but not the only criterion. On DBLP collaboration network and Email-Enron network, Backbone degree algorithm and CNM algorithm has similar performance, but in the same experiment condition, backbone degree algorithm can deal with more data than CNM algorithm. So backbone degree algorithm is superior than CNM algorithm to discover the structure of social networks, and demand for memory is less than the CNM algorithm. But for those communities under 10 vertices backbone degree algorithm gets more greater expansion than CNM algorithm. It is a defect of backbone degree algorithm, we will make up for the defects in the later work.

6. Conclusions

In this paper, we focus on the problem of disjoint community detection in social graphs which is the key tool for understanding the function of the networks and its structure. Many researches in this area are developed and we have discussed their limits in this paper. Our main contributions are three folder. Firstly we propose the community forest model based on these social and biological properties to characterize the structure of real-world large-scale networks. Secondly we mainly define a new metric named backbone degree to measure the strength of the edge and the similarity of vertices and give a new sense definition to community based on expansion. Thirdly we develop a novel algorithm that based on backbone degree and expansion to discover communities from real social networks.

The experiments proved that backbone degree algorithm is superior than CNM algorithm to discover the structure of social networks, and demand for memory is less than the CNM algorithm. Backbone degree algorithm find core backbone edge based backbone degree, then find community based on the trend of expansion decreasing gradually outward from the core backbone. It is starting to computer from local networks, then expand to global networks. First, minimizing the expansion degree of each community, ensuring that expansion degree of each community in the entire network are the smallest, so as to achieve the global optimal. These features can ensure backbone degree algorithm to be extended to parallel computing, so as to deal with large scale social networks.

Backbone degree algorithm is different with all community detection algorithms in Section 2.1, because it is based on a biological and sociological model named Community Forest, and backbone degree algorithm is a simple and direct approach to detect community in networks, it integrated expansion and backbone degree. expansion is used to distinguish the boundaries of communities. Backbone degree integrated network weight and neighborhood overlap, it is very balanced to most of topological structures in networks.

Backbone degree algorithm has good effectiveness and accuracy to social networks, but it only detects disjoint communities in a single-machine environment in undirected networks currently, and there are some little defects such as that expansion is greater than CNM algorithm to those communities under 10 vertices. Our next work is to optimize backbone degree algorithm, and adjust it to suit for detecting overlapping communities from large-scale directed networks in parallel environment.

Acknowledgments

This work is sponsored by National Natural Science Foundation of China (Grant No.: 61175110) and National Basic Research Program of China (973 Program, Grant No.: 2012CB316301). It is also sponsored by the five open platforms foundation of Hebei University of Science and Technology (No.: 2014PT99).

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E.P. (2009). Mixed membership stochastic blockmodels. In *Advances in neural information processing systems* (pp. 33–40).

Arora, S., Rao, S., & Vazirani, U. (2009). Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM), 56*, 5.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*, 066111.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets*. Cambridge University.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*, 7821–7826.

Kannan, R., Vempala, S., & Vetta, A. (2004). On clusterings: Good, bad and spectral. *Journal of the ACM (JACM), 51*, 497–515.

Karypis, G., & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing, 20*, 359–392.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*, 604–632.

Leighton, T., & Rao, S. (1999). Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM (JACM), 46*, 787–832.

Leskovec, J. (2014). <http://snap.stanford.edu>. (supporting website).

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

Leskovec, J., Lang, K. J., & Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on world wide web* (pp. 631–640). ACM.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E, 69*, 066133.

Newman, M. (2009). *Networks: An introduction*. Oxford University Press.

Newman, M. (2013). Spectral methods for network community detection and graph partitioning. In *arXiv preprint* (p. 1307.7729).

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*, 814–818.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 2658–2663.

Rapoport, A. (1953). Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *The Bulletin of Mathematical Biophysics, 15*, 523–533.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, 105*, 1118–1123.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review, 1*, 27–64.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks, 5*, 269–287.

Spielmat, D. A., & Teng, S.-H. (1996). Spectral partitioning works: Planar graphs and finite element meshes. In *37th Annual symposium on foundations of computer science, 1996. Proceedings* (pp. 96–105). IEEE.

Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR), 45*, 43.

Yang, J., & Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD workshop on mining data semantics* (pp. 3). ACM.