Finding overlapping community from social networks based on community forest model

Article *in* Knowledge-Based Systems · July 2016

DOI: 10.1016/j.knosys.2016.07.007

4 authors, including:

Yunfeng xu
College of Information Science and Engineering Hebei University of Science and Technology
23 PUBLICATIONS **262** CITATIONS

SEE PROFILE

Finding overlapping community from social networks based on community forest model

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

[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*

Abstract

Overlapping community detection is the key research work to discover and explore the social networks. A great deal of work has been devoted to detect overlapping communities, but no one can give a clear formula definition of community from the internal structure to the external boundary. More in depth, there are four challenges to existing research works. In this paper, firstly we propose overlapping community forest model and disjoint community forest model based on the community forest model, secondly give a clear formula definition of overlapping community and disjoint community based on the backbone degree and expansion, thirdly propose a novel algorithm to find overlapping communities based on the backbone degree and expansion to resolve the four challenges. This algorithm has better performance than four related algorithms mentioned by this paper in large scale social networks. It works well on American college football, Zachary's Karate Club, Netscience-coauthor, Condensed matter collaborations, LFR etc. data sets.

Keywords: Community detection, social network, expansion, community forest model

*The corresponding author
**† Indicates equal contributions from these authors
Email addresses: hbkd_xyf@hebust.edu.cn (Yunfeng Xu †), xuhua@tsinghua.edu.cn (Hua Xu †), zdwwtx@163.com (Dongwen Zhang), (Yan Zhang)

1. Introduction

Overlapping community detection is the key research work to discover hidden knowledge and structure in social network and other networks, it is also the pre work of propagation analysis and link prediction in complex networks. So it is a very important research. In these research works of Jierui xie et al. (Xie et al., 2013) , Kelley et al. (Kelley et al., 2012) and Reid et al.(Reid et al., 2011), these authors showed that the overlap is a significant feature of many real-world large scale social networks indeed. These overlapping vertices carry more abundant structural information, some of them fill the structural holes between communities, known as the structural hole spanners(Burt, 1994; Lou & Tang, 2013). This overlapping characteristic also happens in other complex networks such as biological metabolic network in a cell, where a metabolic molecule may involve in different metabolic pathways. Communities in a metabolic network might correspond to functional units, cycles, or circuits(metabolic pathways) that perform certain tasks(Newman, 2013). These overlapping communities uncover the general characteristics of the structure in real world networks.

Over the past decade community detection has been applied to many real-world areas such as biological networks, web graphs, VLSI design, social networks, and task scheduling. And recently Wang Xingyuan et al. showed that the communities structure can be used in reality epidemic spreading(Ren & Wang, 2014). Community detection can be categorized into two big classes according to the criterion of whether to allow overlapping: overlapping community detection and disjoint community detection. Disjoint community detection focuses on the division of the boundaries between communities, and the sole ownership of each vertex. Overlapping community detection focuses on the distribution of the membership of each vertex, and the entire global network structure. Disjoint community detection was reviewed and categorized into five research lines, 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 better links between its members than with the remainder of the network(Leskovec et al., 2010). Overlapping community detection 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 et al., 2013), they are investigated based on the consensus that people in a social network are naturally

characterized by multiple community memberships.

Here we mainly introduce the five classes of overlapping community detection algorithm in turn. Firstly, Clique Percolation is based on the assumption that a community consists of overlapping sets of fully connected subgraphs and detects communities by searching for adjacent cliques(Palla et al., 2005). Clique percolation can discover the communities with a variety of density in social networks, and find the structure of the communities. The resolution of community detection is decided by the value of k. But k is a discrete value, so the result of clique percolation is discrete, it means that there are many results with the many k values. Recently, many similar models that coincided with clique percolation have emerged, they inherit the core assumption of Clique Percolation that community gradually expands from the core, such as NDOCD(Z et al., 2016), maximal sub-graph(Cui et al., 2014a), bipartite clustering triangular(Cui et al., 2016) , neighborhood-inflated seed expansion(Whang et al., 2015) and intimate degree(Wang & Li, 2013). Secondly, Line Graph and Link Partitioning are the idea of partitioning links instead of nodes to discover community structure. Although Ahn et al.(Ahn et al.), Evans and Lambiotte(Evans & Lambiotte, 2009; Evans, 2010; Evans & Lambiotte, 2009), Kim and Jeong(Kim & Jeong, 2011) done many valuable works to overlapping community detection based on link partitioning, but these algorithms rely on some ambiguous definitions of community, so there is no guarantee that it provides higher-quality detection than node-based detection does(Fortunato, 2010). Recently Justine Eustace et al. proposed similar algorithms that based on neighborhood(Eustace et al., 2015a)(Eustace et al., 2015b). Thirdly, those algorithms that adopting local expansion metrics and optimization methods are based on growing a natural community(Lancichinetti et al., 2009) or a partial community, most of them rely on some local benefit functions such as topology-potential(Wang et al., 2016; Gan et al., 2009; Han et al., 2011), density function(Kelley & Stephen, 2009; Lancichinetti et al., 2009; Havemann et al., 2010) and some extended modularity metrics(Blondel et al., 2008; Shen et al., 2009a). Fourthly, Fuzzy community detection algorithms adopted the method that quantifying the strength of association between all pairs of nodes and communities(Xie et al., 2013), such as MMSB(Airoldi et al., 2009) and OSBM(Latouche et al., 2011). Fifthly, those agent-based and dynamical algorithms are base on label propagation, information propagation process, Nash local equilibrium, random walks, etc.(Xie et al., 2013).

To sum up, a great deal of work has been devoted to detect overlapping

communities, but the definition of community is still ambiguous(Xie et al., 2013; Z et al., 2016), no one can give a clear formula definition of community from the internal structure to the external boundary. More in depth, there are four challenges to existing research works. Firstly there are no clear boundaries and internal structure for the definition of community. The ambiguity in the definition of community affects the resolution of community detection algorithms. For example, Fortunato et al. find that modularity optimization may fail to identify modules smaller than a scale which depends on the total number L of links of the network and on the degree of interconnectedness of the modules, even in cases where modules are unambiguously defined (Fortunato & Barthelemy, 2007). Secondly they emphasize the optimization of the likelihood function based on consensus, but ignore to discover the boundaries between communities and the community's internal structure. And the community's internal structure and the boundary contain rich information. Thirdly they need complicated iterative and sampling, and thus can not be extended to the current mainstream distributed computing framework. Finally they emphasize physical and mathematical methods to deal with data, while ignoring that social network has a wide range of social and biological characteristics.

How to resolve these four challenges? In this paper, we propose community forest model(CFM for short) algorithm based on extended community forest model(Yunfeng Xu, 2015) to deal with the four challenges in overlapping community detection. Firstly CFM algorithm uses backbone degree and community expansion degree to detect the internal structure and external boundary of community. Community forest model given a clear formula definition of community from the internal structure to the external boundary. Backbone degree measures the strength of the edge and the similarity of vertices(Yunfeng Xu, 2015), and then characterizes the internal structure of the community, we introduce the definition of backbone degree in section 3. Community expansion degree measures the expansion of the community, and the difference of expansion degree measures the change of expansion degree after joining a new vertex to a community, they are used to detect the trends of community expansion degree changes, and then to discover the boundary of the community. So the first challenge is resolved. Secondly CFM algorithm detects community from the core backbone to the boundary, this process is like that finding a tree from forest, it dose not to optimize the likelihood function, so the second challenge is resolved. Thirdly CFM algorithm dose not need iterative and sampling, it can be extended to the

current mainstream distributed computing framework, so the third challenge is resolved. We have implemented CFM algorithm based on hadoop platform, but this paper focuses on algorithms and models, and we will introduce parallel CFM algorithm in subsequent articles. Finally CFM algorithm mines the social and biological characteristics of social network in-depth, and calculates simple and effective measures such as expansion and backbone degree, the process of community detection does not need complicated physical and mathematical methods, so the fourth challenge is resolved.

Why we extend community forest model to resolve the overlapping community detection problem in this paper? Compared with the other related representative models that using clique percolation(Palla et al., 2005), neighborhood(Eustace et al., 2015a)(Eustace et al., 2015b), maximal subgraph(Cui et al., 2014a), intimate degree(Wang & Li, 2013), community forest model can give a global perspective to discover social networks, which is a simple and effective social network model based on the characteristics of biology and sociology. The social and biological characteristics of community and network refer to the characteristics that got from the study of microscopic feature of the specific network(Yunfeng Xu, 2015). For example, neighborhood overlap (Easley & Kleinberg, 2010), authority weight, hub weight (Kleinberg, 1999), growth form and structure, succession(Connell & Slatyer, 1977) and so on. Based on these characteristics, we proposed a new definition of community, a global measure named backbone degree and community forest model(Yunfeng Xu, 2015), but community forest model were only applied to discover disjoint communities. However there are usually two kinds of relationships between communities in large-scale social networks: disjoint and overlapping. There are clear boundaries between disjoint communities and their adjacent communities, and the boundaries between overlapping communities and their adjacent communities are blurred. And in large scale real social networks, disjoint and overlapping communities are coexisting—some communities with clear boundary and other communities with blurred boundary. Using community forest model to identify the boundaries of overlapping communities may loss some information, for example the membership of overlapping vertices. So we must extend the community forest model to overlapping community forest model and disjoint community forest model, and develop CFM algorithm to find overlapping communities from social networks based on extended community forest model.

In this paper, Firstly we extend the community forest model to overlapping community forest model and disjoint community model based on these

social and biological properties. Secondly we mainly propose the definition of overlapping community and disjoint community based on expansion and backbone degree. Thirdly we develop CFM algorithm based on backbone degree and expansion to discover overlapping 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 survey preliminary concepts about community detection and discuss the physical meaning of these concepts. In Section4, we describe the model and problem formalization of this paper. In Section 5, we systematically develop CFM algorithm. Section 6 is experiment study and Section 7 concludes this study.

2. Related Work

In this section, we survey these research works that very relevant to our work: Clique Percolation(Palla et al., 2005), Local Expansion and Optimization, Line Graph and Link Partitioning, MMSB model(Airoldi et al., 2009) , Louvain mothed, backbone degree algorithm(Yunfeng Xu, 2015), F-measure metric(Artiles et al., 2007; Banerjee et al., 2005) and extended modularity Q_m (Li et al., 2014) .

We get some inspiration from clique percolation. Clique percolation is based on the assumption that a community consists of overlapping sets of fully connected subgraphs and detects communities by searching for adjacent cliques. Clique percolation method traverses k-clique in the social networks, when k=3, it coincides with the neighborhood overlap in the backbone degree concept. And Yang et al. found that when k=4, CPM can get more good result than k equals other values (Yang & Leskovec, 2013). Backbone degree integrates the triadic closure, neighborhood overlap and network weight(Yunfeng Xu, 2015), it is similar with k-clique. However backbone degree quantifies the k value of k-clique smoothly based on neighborhood overlap and network weight. In recent years, there are many related studies that similar with clique percolation algorithm, such as EM-BOAD(Li et al., 2013a), Cui's algorithms(Cui & Wang, 2014; Cui et al., 2016), Li's algorithm(Li et al., 2014), NDOCD(Z et al., 2016), NISE(Whang et al., 2016). All of these similar algorithms inherit the core assumption of clique percolation that community gradually expands from the core.

Line Graph and Link Partitioning are the idea of partitioning links instead of nodes to discover community structure(Xie et al., 2013). Ahn et al.(Ahn

et al.), Evans and Lambiotte(Evans & Lambiotte, 2009; Evans, 2010; Evans & Lambiotte, 2009), Kim and Jeong(Kim & Jeong, 2011) done many valuable works to overlapping community detection based on link partitioning, these ideas coincide with the backbone degree algorithm(Yunfeng Xu, 2015) partly that detecting community based on the backbone degree of edges.

We inspired from local expansion and optimization algorithms, then pulled the expansion to detect the boundary of community. Those algorithms that adopting local expansion metrics and optimization methods are based on growing a natural community(Lancichinetti et al., 2009) or a partial community, NLA algorithm(Wang et al., 2016) is such class of algorithm in recent years. Their strategy of detecting community coincides with our algorithm in this paper. Yunfeng Xu et al. found that expansion can detect the clear boundary of community(Yunfeng Xu, 2015). In this paper, we will refine the definition of the expansion based on backbone degree, then extend disjoint community detection to overlapping community detection.

MMSB model(Airoldi et al., 2009) given some inspirations to us. The similarity between vertices can be quantified based on the probability of mixed membership, so the similarity between vertex and community also can be quantified based on the probability of mixed membership. We propose the definition of probability membership based on backbone degree and mixed membership in this paper, this definition can give perfect boundary to overlapping community and disjoint community.

Louvain method is a greedy optimization method that attempts to optimize the "modularity" of a partition of the network. The optimization is performed in two steps(Blondel et al., 2008). First, the method looks for "small" communities by optimizing modularity locally. Second, it aggregates vertices belonging to the same community and builds a new network whose vertices are the communities. These steps are repeated iteratively until a maximum of modularity is attained and a hierarchy of communities is produced. The first step of louvain method is similar with finding k-clique in CPM algorithm. The modularity metric that adopted by louvain method has similar function with the expansion in CFM algorithm, but CFM extended expansion based on backbone degree, and adopted several metrics that proposed in this paper.

We inspired from these concepts of neighborhood overlap, authority, structural hole, etc. in Kleinberg's book (Kleinberg, 1999). kleinberg mined these characteristics of social networks deeply, the problem of community detection can be solved by some simple and efficient solutions based on these character-

istics of social networks. And these concepts give the possibility of analyzing social networks without the shackles of the methods of pure mathematics and physics.

We adopt F-measure to evaluate overlapping community detection in LFR data set. Banerjee et al. adopted the same evaluation method(Banerjee et al., 2005). F-measure is an outcome of information retrieval, which is the harmonic mean of both versions of the purity(Artiles et al., 2007). Amig et al. proved that only BCubed(Bagga & Baldwin, 1998) satisfies all formal constraints through their analysis of a wide range of metrics, and extended the analysis to the problem of text overlapping clustering(Amig et al., 2009). But we are not sure that BCubed metric is perfect to overlapping community detection, because compared with the problem of text overlapping clustering , the problem of overlapping community detection has its own unique characteristics, for example these social and biological characteristics that pointed out in this paper, we will use BCubed metric(Amig et al., 2009) in our future work for evaluation. And we believe that only F-measure is not enough, Amig et al. proved that evaluation by set matching, metrics based on counting pairs, metrics based on entropy and evaluation metrics based on edit distance do not satisfy all formal constraints. F-measure, Mutual Information, Purity and Inverse Purity etc are all included in these metrics.

We adopt Q_m to evaluate these real world social networks without standard partition results. Q_m is proposed by Junqiu Li et al.(Li et al., 2014). It is adapted from an extended measure Q_c proposed by Shen et al.(Shen et al., 2009b). Q_c is based on a maximal clique view of the original network. Shen et al. proved that the optimization of Q_m on the original network is equivalent to the optimization of the modularity on the maximal clique network. Q_c and Q_m are all the extension of modularity Q(Newman & Girvan, 2004). We consider that Q_m is a reasonable evaluation metric for large real world social networks without standard partition results.

Other related research lines are shown below. Wang xingyuan et al. proposed an algorithm that using the core-vertex and intimate degree, this algorithm finds the core-vertices as the initial community, then expands it using intimate degree function during extracting community structure from the given network(Wang & Li, 2013). This algorithm and backbone degree algorithm(Yunfeng Xu, 2015) are based on the similar hypothesis that communities are expanding from core-vertices or core-backbone edges. Li, Yakun et al. proposed a novel definition for community and an efficient community detection algorithm which takes advantage of additive topological and other

constrains to discover communities in arbitrary shape based on the feedback(Li et al., 2013b), compared with CFM algorithm, this work ignores the depiction of the internal structure of the community. Cui et al proposed BASH and ACC algorithm, these two algorithms are similar, they extract all the maximal sub-graphs firstly, and then merge two maximal sub-graphs having the key pair-vertices into a new sub-graph using belonging degree functions or the clustering coefficient, and extend the quality function of modularity to evaluate these results(Cui et al., 2014a)(Cui et al., 2014b). Zhongying Zhao et al. proposed a topic oriented community detection approach which combines both social objects clustering and link analysis, and they used a subspace clustering algorithm to group all the social objects into topics(Zhao et al., 2012). Justine Eustace et al. proposed an algorithm to approximate web communities from the topic related web pages, and introduced a new framework that reduces the impact of noise-links in detecting a community using topic-related web pages(Eustace et al., 2014), then Justine Eustace et al. proposed two algorithms that using the overlapping neighborhood ratio(Eustace et al., 2015b) and local community neighborhood ratio function(Eustace et al., 2015a).

3. Preliminaries

In this section, we survey preliminary concepts of community detection, and discuss the physical meaning of these concepts.

Community: Definition 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 et al., 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. We propose the definition of community in our previous research work based on expansion and backbone degree(Yunfeng Xu, 2015), this notion can visualize the notion of community, which means that the notion can give clear boundary and internal structure.

Disjoint Community and Overlapping community: Definitions of disjoint community and overlapping community are too common, so no clear definition had been given to them. Currently the consensus of disjoint community: any vertex in the network belongs the exclusive community. The consensus of overlapping community: any vertex in the network can belong more than one community, it is well-understood that any people in a real social network may have multiple community memberships(Xie et al., 2013). In social networks, there are always some active vertices with multiple identities and multiple social roles, they play different roles in multiple communities, and they can meet the characteristics of the community in the process of dividing the community, such as vertex 10, 9, 31 in the figure 1. According to the principle of triangle closure, these vertices can be divided into the community with the closure relation. But there is also the link between each vertex to its associated community is equal, in accordance with the principle of community division, it can not be divided into any community, such as vertices are called structural hole spanners. If these structural hole spanners strengthen the connection with each of the connected communities, such as increasing the number of friends, they will cause communities to overlap, and even the integration of communities. So we think that overlapping community and disjoint community are relative, coexisting and transformed to each other in real social networks. So far, all of the concepts of overlapping community and disjoint community are based on a vertex that allowed to belong to multiple communities. There is no definition that defines the overlapping community from the perspective of community and quantifies the degree of overlap, this paper attempts to propose a model that can define the overlapping community from the perspective of community and quantify the degree of community overlap. We will give clear formula definition to disjoint community too.

backbone degree: Kleinberg et al. (Easley & Kleinberg, 2010) defined closure, structural holes, weak and strong link, bridge, shortcut, neighborhood overlap, etc. The neighborhood overlap characterizes the strength of an edge. On the basis of these concepts, we putted forward the metric backbone degree to characterizes the strength of the links between vertices and communities(Yunfeng Xu, 2015). A backbone consists of an edge and two

vertices that connecting to the edge. Backbone degree measure three factors: an edge and two vertices. The edges like the sections of the bamboo. Every section consists of two joints and a bar. The relation is the bar, 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. Backbone degree is defined in definition 1.

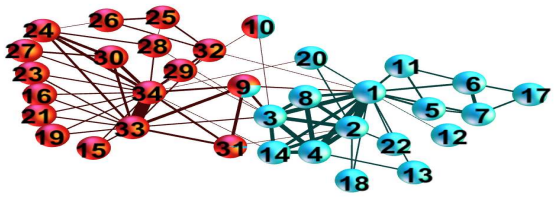

Figure 1: The result of applying CFM algorithm to Zachary Karate Club network.

Expansion: Ravi 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 vertices 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.

Resolution: Fortunato et al. found that modularity optimization may fail to identify modules smaller than a scale which depends on the total number L of links of the network and on the degree of interconnectedness of the modules, even in cases where modules are unambiguously defined (Fortunato & Barthelemy, 2007). Resolution determines the depth and breadth of social network mining, and determines the extent of the characterization to the social network structure. However, is there a metric to control the resolution? We believe that controlling the threshold of some metrics can determine the resolution of community detection, such as expansion.

The probability that vertex i belongs to community C:

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

In social networks, there are always some active vertices with multiple identities and multiple social roles, they play different roles in multiple communities. Most of these vertices can be divided a mainly community based on the above probability. But if we do not focus on the probability of any vertex, but the expansion of community, can we discover the overlapping communities in the social network? The result is true. We find that if we do not collect the vertices based on the probability in backbone degree algorithm(Yunfeng Xu, 2015), we can get these overlapping communities in the social network. Why expansion divide overlapping communities from the social network, but dose not divide disjoint communities? Because expansion characterizes the community from the whole community perspective, but not anyone vertex perspective, it ignores the ownership of single vertex. But this weakness of expansion leads it to discover overlapping communities. In this paper we extend expansion metric based on backbone degree and membership probability to find disjoint community. $P(i \in C)$ coincides with MD in NDOCD(Z et al., 2016) and absorbing degree in EM-BOAD(Li et al., 2013a).

Classification of overlapping communities: Through reading the papers and observing the results of community detection, we find that overlapping communities can be divided into 4 classes based on the perspective of community: none overlap, boundary overlap, partial overlap, matryoshka doll overlap(hierarchical structure). In real world large scale social networks, these 4 classes overlap may coexist, so a good algorithm should be able to discover both overlapping communities and hierarchies between them(Lancichinetti et al., 2009). We find that CFM algorithm can find matryoshka doll overlap and other 3 classes overlap by refining the expansion of community based on backbone degree, we do some experiments on zachary karate club and American college football data sets to prove it. These experiments are showed in section 6.

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 to 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 weight of vertices can be considered $x_j = \frac{d_j}{2m}$ then

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

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. Neighborhood overlap is similar with the intimate degree(Cui & Wang, 2014) between two vertices u and v.

$$NO_{uv} = \begin{cases} \dfrac{|NB_v \bigcap NB_u|}{|NB_v \bigcup NB_u| - 2}, & \text{there is an edge between } u \text{ and } v; \\ 0, & \text{there isn't an edge between } u \text{ and } v. \end{cases}$$

Extended F-measure: F-measure is an outcome of information retrieval, which is the harmonic mean of both versions of the purity(Artiles et al., 2007). One version of mean is named recall, other is named precision.

$$F(X,Y) = \frac{2 * Pur(X,Y) * Pur(Y,X)}{Pur(X,Y) + Pur(Y,X)}$$

Let $Y = (C_1, C_2, C_3..., C_k)$, $\bigcup_{(i..k)} C_i = V$ is a original community division of data set, let $X = (C'_1, C'_2, C'_3..., C'_k)$, $\bigcup_{(i..k)} C'_i = V$ is a community division to be evaluated of data set. Let (u, v) is a pair of vertices, it is different with the edge definition in section 4. The definitions of recall, precision and F-measure are showed below.

$$recall = \frac{|\{(u,v) : u \in C'_i, v \in C'_i\} \bigcap \{(u,v) : u \in C_i, v \in C_i\}|}{|\{(u,v) : u \in C_i, v \in C_i\}|}$$

$$precision = \frac{|\{(u,v) : u \in C'_i, v \in C'_i\} \bigcap \{(u,v) : u \in C_i, v \in C_i\}|}{|\{(u,v) : u \in C'_i, v \in C'_i\}|}$$

$$F - measure = \frac{2 * recall * precision}{recall + precision}$$

Extended modularity Q_m : The formula of Q_m is as follows.

$$Q_m = \frac{1}{2m} \sum_{c \in P} \sum_{uv} \beta_{uc} \beta_{vc} (A_{uv} - \frac{k_u k_v}{2m}),$$

where P is a partition of network, c is any one community in P,

$$\beta_{uc} = \frac{1}{N_{u \in c}} \frac{n_{uc}}{k_u},$$

and $N_{u \in c}$ is the count of vertex u belonging to communities, n_{uc} is the count of edges adjacent to vertex u in the community, $k_u = \sum_v A_{uv}$ is the sum of the edges connected to vertex u. The definition of Q_m satisfies the following conditions

$$0 \leq \beta_{uc} \leq 1, \forall c \in P, u \in V.$$

If vertex u belongs to only one community c, β_{uc} is equal to 1. Then Q_m is consistent with the definition of the modularity Q.

4. Model and Problem Formalization

In this section we introduce overlapping and disjoint community forest model and community detection problem description firstly, then give the problem formalization of overlapping community detection.

4.1. Model and Problem Description

In large-scale social networks, overlapping communities and disjoint communities are coexistent, and a social network contains more overlapping communities, the social network is more active. What causes communities to change from disjoint communities to overlapping communities? We think there are two reasons: Firstly the strengthen of relationship between the core vertices in communities. Secondly the structural hole spanners strength the relationship with their connected communities. The two reasons have same characteristic that all these core vertices and structural hole spanners have mixed membership. The mixed membership can not be detected before dividing the social network completely.

Airoldi et al. proposed the mixed membership stochastic block models(MMSB) to characterize and detect the mixed membership of vertices(Airoldi et al., 2009), and developed a general variational inference algorithm for fast

approximate posterior inference. MMSB is a class of latent variable models. We proposed the community forest model to characterize the social networks in our last paper(Yunfeng Xu, 2015), but we do not give clear formula definition to disjoint community and overlapping community in last paper. Clear formula definition of community is the key process to discover overlapping community from social network, and it is the baseline of realizing community detection algorithm with program. We will give the clear formula definition of disjoint and overlapping community in this paper. Community forest model given the definition that community is a set with some vertices that expand outward from the core backbone according the backbone degree gradually, and the expansion diminishes gradually, until the difference of expansion is minimum. The expansion metric characterizes the community from the whole community perspective, but not single vertex perspective, it ignores mixed membership of a single vertex. But this weakness of expansion leads it to discover overlapping communities. In community forest model, the mixed memberships of vertices are shown after dividing the social network completely, this point is same as CPM(Palla et al., 2005).

The difference between the disjoint and overlapping community detection algorithm is that whether the vertex has mixed membership. And in the real social network, there are a large number of vertices with mixed membership. Many disjoint community detection algorithms adopt the spectrum method and modularity metric to identify the sole ownership community of the vertices, this kind of approach is very elegant, clear and useful to some application scenarios, but leads to the loss of hidden structural information. Many overlapping community detection algorithms adopt probabilistic model and statistical model to approximate the mixed membership of vertices. For example MMSB, airoldi et al. adopted variational inference algorithm to approximate the posterior inference of the mixed membership, which is a kind of advanced structural information mining model, but it needs many iterations, so the computing cost of it is very high.

In this paper, we propose a probability membership metric based on backbone degree, which can be used smoothly from disjoint community detection to overlapping community detection. Based on this metric, we can clearly define overlapping community forest and disjoint community forest. The overlapping community forest is a set of community that conforms to the basic characteristics of the community definition in community forest model, and does not pay attention to the mixed membership of the specific individual vertex. Disjoint community forest is a set of community that conforms

to the basic characteristics of the community definition in community forest model, and each vertex belongs to the community with maxim probability membership.

According to the community forest model, the process of overlapping community detection can be defined like that: finding the core backbone of each community and looking for the boundary of each community(Yunfeng Xu, 2015). In CFM, the process of finding the core backbone of each community likes that finding the core trunk of each tree in the forest. So the process needs a convenient and practical metric to measure the edges in social network, the result proved that backbone degree is the right metric in our last paper(Yunfeng Xu, 2015). The process of looking for the boundary of each community likes that finding the boundary of the tree in the forest. We found that the trunk gradually became thin from tree backbone to leaves, and became thick from leaves to tree backbone. This trend can be described by the expansion metric and backbone degree(Yunfeng Xu, 2015). We refine the expansion metric based on the backbone degree in this paper, and then improve the resolution of overlapping community detection algorithm. By mining multiple core backbone, and then refining the boundary resolution, we can further detect hidden multiple community in the network.

4.2. Problem Formalization

Given an undirected graph $G(V, E)$ with $|V|$ vertices, $|E|$ edges and K communities. Let $n = |V|$, $m = |E|$, C_k is a community in G, $k = 1, 2, ..., K$. $|C_k|$ is the number of vertices in C_k, for simple C can represent C_k. Let $E_C = \{(u, v) \in E : u \in C, v \in C\}, |E_C|$ is the number of edges in C. Let $C_{BE} = \{(u, v) \in E : u \in C, v \notin C\}$, $|C_{BE}|$ 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 vertex set of vertex u. Let NB_C be the neighborhood vertex set of community C, $NB_C = \{v : (u, v) \in E, u \in C, v \notin C\}$.

Definition 1. ***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(Yunfeng Xu, 2015).

Definition 2. ***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. S represents sum of backbone degree, C represents the community C. D_{uv} is the backbone degree of the edge (u, v).*

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

Definition 3. ***Sum of backbone degree to the edges of vertex v not point to community C:*** *this metric can measure the distance between vertex v and other communities without community C. S represents sum of backbone degree, N represents that dose not point to community C. C represents the community C.*

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

Definition 4. ***Probability Membership:****The probability that the vertex v belongs to the community C.*

$$P(v \in C) = \frac{SC_v}{SC_v + SNC_v}$$

Definition 5. ***Community Expansion Degree:*** *this metric measures the number of edges that point outside the community.*

$$EX_C = \frac{|C_{BE}|}{|C|}$$

Definition 6. ***Sum of backbone degree to the edges of*** C_{BE}***:*** *this metric can measure the distance between vertex v and other communities without community C.*

$$SBD_{C_{BE}} = \sum_{(u,v) \in C_{BE}}^{C_{BE}} D_{uv}$$

Definition 7. ***Community Expansion Degree based backbone degree:*** *this metric measures the expansion of the community based on the backbone degree. $EXBD_C$ can quantify the expansion of the community C more accurately than EX_C.*

$$EXBD_C = \frac{SBD_{C_{BE}}}{|C|}$$

Definition 8. ***Overlapping Community:*** *This paper adopts the definition of community in our last paper(Yunfeng Xu, 2015). 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.*

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

We refine the metric expansion based on backbone degree, we get more detailed overlapping community definition. Limiting conditions changed from EX_C to $EXBD_C$.

$$OC = \{u : u \in C, v \notin C, (u, v) \in E, EXBD_{\{C \bigcup v\}} > EXBD_C\}$$

Definition 9. ***Disjoint Community:*** *This paper gives a new definition of disjoint 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, and each vertex belongs to the community with maxim probability membership.*

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

$$P(u \in C) = arg \max_{C_k} P(u \in C_k)\}$$

If applications need more detailed communities, limiting conditions can be changed from from EX_C to $EXBD_C$.

$$DC = \{u : u \in C, v \notin C, (u, v) \in E, EXBD_{\{C \bigcup v\}} > EXBD_C,$$

$$P(u \in C) = arg \max_{C_k} P(u \in C_k)\}$$

Definition 10. ***Overlapping Community Forest:*** *The overlapping communities set in G can be defined as community forest. Overlapping community forest is a set of community that conforms to the basic characteristics of the community definition in community forest model, and does not pay attention to the mixed membership of the specific individual vertex.*

$$OCF = \{C : C \subseteq V, v \notin C, EX_{\{C \bigcup v\}} > EX_C\}$$

if we refine the metric expansion based on backbone degree, we will get more detailed overlapping community forest.

$$OCF = \{C : C \subseteq V, v \notin C, EXBD_{\{C \bigcup v\}} > EXBD_C\}$$

Definition 11. ***Disjoint Community Forest:*** *The communities set in G can be defined as community forest. Disjoint Community Forest is a set of community that conforms to the basic characteristics of the community definition in CFM, and each vertex belongs to the community with maxim probability membership. $EXBD_C$ can be changed to EX_C when applications need.*

$$DCF = \{C : C \subseteq V, v \notin C, u \in C, EXBD_{\{C \bigcup v\}} > EXBD_C,$$

$$P(u \in C) = arg \max_{C_k} P(u \in C_k)\}$$

Definition 12. ***The Difference of Expansion Degree:*** *The change of expansion degree after joining a new vertex i to community C.*

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

5. Algorithm

In this section, we introduce the design of CFM algorithm firstly, secondly use pseudocode to describe the algorithm framework, thirdly analyze the time complexity of CFM algorithm, finally give the algorithm analysis and comparison to CFM algorithm and other ten related state-of-the-art algorithms.

5.1. Algorithm Design

According to the community forest model, the process of CFM algorithm can be defined like that: finding the core backbone of each community and looking for the boundary of each community. The detailed process can be described as that: Firstly CFM algorithm 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. Secondly CFM algorithm creates an empty community, and selects the backbone with the maximum backbone degree in the list as the initial backbone to the current community, then adds the neighboring vertex with the maximum SC in turn. The neighboring vertex is in the neighbor set of the current community. SC is defined in definition 2. If the expansion becomes smaller after adding a new neighboring vertex to the current community, then CFM algorithm continually adds the neighboring vertex with the maximum SC. Else CFM algorithm adds the neighboring vertex to the boundary set of the

current community, and continues to find the vertex with the maximum SC that connected with the current community until there is no longer eligible vertex in the neighbor set of the current community. At this point a new community is divided completely right now. Thirdly CFM algorithm repeats the second step, divides 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. Finally CFM algorithm collects these vertices that divided into no community and multi-community based on SC, this process sorts the probability membership of these vertices, then identifies the mixed membership of these vertices. Probability membership is defined in definition 4. Mixed membership is point out in MMSB model(Airoldi et al., 2009). The pseudo code of CFM algorithm is shown in algorithm 1.

The parameter 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 probability membership 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 backbone degree of core backbone is more than 0.3, then let $f = 0.3$. When the social network is small, w and f can be zero.

Why CFM algorithm uses the expansion degree to determine the ownership of vertex rather than the probability membership? Because the probability membership of vertex can not be calculated before dividing core communities completely. Community forest model is based on the hypothesis that the expansion of community diminishes gradually from core backbone to boundary. We proved that this hypothesis is simple and correct through experiments in our last paper(Yunfeng Xu, 2015). And we propose two expansion definitions in definition 5 and 7, definition 7 is more detailed expansion than definition 5.

Why CFM algorithm starts to divide the community from the core backbone? If we consider the community as a tree in the forest, based on the assumption that the community must expand from the core backbone. This is a simple consensus, and we proved this question in our last paper(Yunfeng Xu, 2015).

5.2. Framework of CFM 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. CFM algorithm implementation is shown in Algorithm 1. The expansion can be changed from definition 5 to definition 7 based on the needs of applications.

5.3. Algorithm Time Complexity

CFM algorithm uses merge sort to sort the backbone list, it runs in time $O(mlogm)$, and the process of discovering community runs in time $O(n+m)$, so our algorithm runs in time $O(mlogm+n+m)$ 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, $O(mlogm)$ will fall sharply too, so our algorithm runs in time $O(n + m)$ approximately. We analyzed backbone degree of five data set. Table 3 is the edges with maximum backbone degree. Table 4 is the count of edges with $f \geq 0.2$ and $f \geq 0.3$. We found that the maximum 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 4.

5.4. Algorithm Analysis and Comparison

CFM algorithm inherits the core assumption of Clique Percolation that community gradually expands from the core. In recent years, there are many related studies that similar with CFM algorithm, such as EM-BOAD(Li et al., 2013a), Cui's algorithms(Cui & Wang, 2014; Cui et al., 2016), Li's algorithm(Li et al., 2014), NDOCD(Z et al., 2016), NISE(Whang et al., 2016). All of these similar algorithms inherit the core assumption of Clique Percolation that community gradually expands from the core, and adopt the community detection strategy that finding seed community or core community firstly then expanding them based on some metrics such as joint strength(Z et al., 2016), intimate degree(Wang & Li, 2013), absorbing degree(Li et al., 2013a) and membership degree(Z et al., 2016), although they chose and expand the core of community in different ways. And Cui's algorithms(Cui & Wang, 2014; Cui et al., 2016) applied this kind of community detection strategy

Algorithm 1 CFM Algorithm Implementation based on community forest model

Data: An undirected $G(V, E)$. **Result:** The community set OCF in G.
begin
1 $NL \Leftarrow V$, $BL \Leftarrow$ *edges with backbone degree* $\geq f$ *in* E, $OCF \Leftarrow null$, $i \Leftarrow 0$. Sort BL according to descending order, $index_{BL} \Leftarrow 0$.
2 Get a backbone b from BL according to $index_{BL}$, $index_{BL}++$, get vertex u and v from b, note the backbone degree of b as BD_b, note the size of NL as nl.
3 **while** $BD_b \geq f$ and $nl \geq w$ **do**
4 **if** $u \in NL$ and $v \in NL$ **then**
5 $C_i \Leftarrow \{u, v\}$
6 $EX_{C\_}PRE \Leftarrow$ the EX 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 OCF; $i++$; goto step2.
10 **else**
11 find the nearest vertex nv from $\{NB_{C_i} - BV_{C_i}\}$ based on SC_{nv}, add vertice nv to C_i, calculate the EX of C_i and note it as $EX_{C\_}cur$.
12 **if** $(EX_{C\_}cur - EX_{C\_}PRE < 0)$, **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 OCF, $i++$, 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 these vertices that divided into no community and multi-community based on SC(the short form of definition 2).
27 return OCF.
end

to bipartite network, EM-BOAD(Li et al., 2013a) applied this kind of community detection strategy to weighted network. Compared to these similar algorithms, CFM algorithm depicts the internal structure and external boundary of the community based on the backbone degree and EX/EXBD in depth, while other algorithms only focus on how to detect communities in the networks.

CFM algorithm detects the boundary of community through the infinite approximation to the minimum difference of expansion degree, its detecting communities strategy is similar with those algorithms that adopting local expansion metrics and optimization methods. But CFM algorithm sorts all edges in network based on backbone degree and selects core community in the sorting edge list, this strategy guarantees the global optimization. Backbone degree is a more simple and effective global metric than other metrics such as topology-potential(Wang et al., 2016; Gan et al., 2009; Han et al., 2011) and CNFV(Lei et al., 2013; Z et al., 2016) in social networks. Because backbone degree integrated EX and neighborhood overlap can choose the core of the community more accurately , and avoid those vertices such as structural hole spanners(Yunfeng Xu, 2015). We consider that structural hole spanners and real core vertex of community all can get high value of topology-potential and CNFV in sparse social networks, this may affect the accuracy of NLA algorithm(Wang et al., 2016), NDOCD(Z et al., 2016) algorithm and and similar algorithms that selecting vertices as seed communities.

Through the above analysis, we find that CFM algorithm is similar with first class and third class mentioned by Xie et al.(Xie et al., 2013). In table 1, we compare the time complexity of ten related state-of-the-art algorithms with CFM algorithm, and show the category of all eleven algorithms. CPM belongs to the first class. NDOCD(Z et al., 2016), EM-BOAD(Li et al., 2013a), Cui's algorithm A(Cui & Wang, 2014), Cui's algorithm B(Cui et al., 2016), NISE(Whang et al., 2016) and Li's algorithm(Li et al., 2014) are similar with first class and third class , NLA(Wang et al., 2016) and louvain method belong to the third class, MMSB(Airoldi et al., 2009) belongs to the fourth class mentioned by Xie et al.(Xie et al., 2013).

CPM employed many community definitions by allowing incomplete k-cliques or decreasing k, so the time complexity is NP-complete. Louvain method is a greedy optimization method, the optimization includes two steps. First, the method looks for "small" communities by optimizing modularity locally. Second, it aggregates vertices belonging to the same community and builds a new network whose vertices are the communities. These steps are

repeated iteratively until a maximum of modularity is attained and a hierarchy of communities is produced. The method seems to run in time $O(nlogn)$ with most of the computational effort spent on the optimization at the first level. Exact modularity optimization is NP-hard. So the time complexity of louvain is $O(nlogn)$ on the first level optimization, and NP-hard on exact modularity optimization. MMSB algorithm interleaves subsampling from the network and updating an estimate its communities(Gopalan & Blei, 2013), so the time complexity of it can not be estimated clearly. NDOCD algorithm includes three steps: seed selection,seed expansion and network decomposition(Z et al., 2016). The time complexity of seed selection is approximately equal to O(n), the time complexity of seed expansion is approximately equal to $O(m)$, the time complexity of network decomposition is approximately equal to $O(m)$, so the total time complexity of NDOCD is approximately equal to $O(n + 2m)$, and then approximately equal to $O(n + m)$. The time complexity of NLA is approximately equal to $O(n^2)$ (Wang et al., 2016). The worst time complexity of EM-BOAD is upper bound at most $O(n^2)$(Li et al., 2013a). The worst time complexity of Cui's algorithm A is at most $O((m + n)^2)$(Cui & Wang, 2014). Cui's algorithm B (Cui et al., 2016) incudes one step that extracting all the maximal sub-graphs, this step is same with CPM, so the time complexity of Cui's algorithm B is approximately NP-complete. NISE algorithm includes four steps: Filtering, Seeding, Seed expansion and Propagation(Whang et al., 2016), the time complexity of filtering phase is $O(n+m)$. The remaining three steps are related to the number of seeds, and the complexity is difficult to estimate based on n and m. But the remaining three steps visited all the vertices and edges at least once, because NISE can find overlapping vertices. So the time complexity of NISE is approximately equal to $O(n + m)$ (Whang et al., 2016). Li's algorithm extracts all maximal cliques in networks, so its time complexity is same with CPM. CFM algorithm runs in time $O(n + m)$ approximately.

Table 1: The time complexity and category summary of CPM, louvain method, MMSB, NDOCD, NLA, EM-BOAD, Cui's algorithm A(Cui & Wang, 2014), Cui's algorithm B(Cui et al., 2016), NISE(Whang et al., 2016), Li's algorithm and CFM.

Algorithms	Time complexity	Category
CPM	NP-complete	1
Louvain method	$O(nlogn)$ or NP-hard	3
MMSB	Can not be estimated clearly	4
NDOCD	$O(n+m)$ approximately	1,3
NLA	$O(n^2)$	3
EM-BOAD	Upper bound at most $O(n^2)$	1,3
Cui's algorithm A	At most $O((m+n)^2)$	1,3
Cui's algorithm B	NP-complete	1,3
NISE	$O(n+m)$ approximately	1,3
Li's algorithm	NP-complete	1,3
CFM	$O(n+m)$ approximately	1,3

Through the analysis of the above algorithms, we found the following 3 rules: Firstly, the time complexity of those algorithms that based on finding the maxim sub-graph is determined by the distribution of the k in k-clique, so this kind of algorithm is NP-complete, such as CPM, Cui's algorithm B and Li's algorithm. Secondly, the time complexity of those algorithms that based on optimizing the local metric of the community is $O(n+m)$ approximately, such as NDOCD, NISE and CFM. Thirdly, the time complexity of those algorithms that based on optimizing the global metric of the community and generating model is NP-hard or not be estimated clearly, such as louvain method, NLA and MMSB. Because Cui's algorithm A deals with bipartite network, and EM-BOAD deals with weighted network, so they have a relatively high time complexity.

Time complexity does not represent the actual running time of the algorithm, and some algorithms are very sensitive to the structure and scale of the network. And the time complexity of MMSB, CPM and louvain is not clear, so we do the test for run time comparison in table 8. It is hard to say which algorithm is better in time complexity, because the application requirements determine how to use these algorithms.

CFM algorithm is based on backbone degree and overlapping communi-

ty forest model, and it is scalable and can be applied to large-scale social networks. CFM algorithm can discover networks in different depth based on backbone degree threshold f and parameter w, this made it very flexible.

6. Experiments

In this section, we study the effectiveness and accuracy of CFM algorithm and compare it with CPM (Palla et al., 2005) algorithm, MMSB algorithm(Gopalan & Blei, 2013) and Louvain algorithm(Blondel et al., 2008), these three algorithms are all relevant to CFM algorithm. And we study the performance of CPM algorithm, MMSB algorithm, Louvain algorithm, Li's algorithm(Li et al., 2014) and CFM algorithm based extended modularity Q_m in large real world social networks. We get CFinder from the website www.cfinder.org, CFinder is a free software for finding and visualizing overlapping dense groups of nodes in networks, based on the Clique Percolation Method (CPM) of Palla et. al., Nature 435, 814-818 (2005). We get SVINET from website http://github.com/premgopalan/svinet, SVINET is the implementation of MMSB algorithm. We get Louvain method from http://sites.google.com/site/findcommunities/, the Louvain method is a simple, efficient and easy-to-implement method for identifying communities in large networks. We can not get all demo programs of other seven algorithms that mentioned in section Algorithm Analysis and Comparison, so the comparison of performance is not complete. But We believe that the existing experiments can prove that CFM is an algorithm with satisfactory performance.

6.1. Experimental Design and Overview

We employ four evaluation methods to test CPM, MMSB, CFM and Louvain algorithms, and we compared these four algorithms with Li's algorithm.(Li et al., 2014) through citing the experiments in their paper: Firstly we study the effectiveness and accuracy of CPM, MMSB, CFM and Louvain algorithms based on American College Football and Karate Club data sets. Secondly we adopt the metric Q_m to evaluate five algorithms based on American College Football and Karate Club, Netscience-coauthor and Condensed matter collaborations 2003 data sets. Thirdly we adopt the metric F-measure to evaluate three algorithms based on LFR benchmark data set. Finally we compare the run time of CPM, MMSB, CFM and Louvain algorithms based on LFR benchmark data set. American College Football and Karate Club

are small standard data sets, they can be visualized to show the effectiveness and accuracy of CFM algorithm and other algorithm. Netscience-coauthor and Condensed matter collaborations 2003 are large real-world networks, but they have no standard partitioning results. So we adopt Q_m proposed by Junqiu Li et al.(Li et al., 2014) to evaluate the performance of CPM, MMSB, CFM and Louvain algorithms in large real-world networks. Because we have no demo of Li's algorithm, we have not calculated the Q_m of LFR benchmark data set to compare with Li's algorithm.

In these four evaluation methods, we set parameters w and f to 0 for accuracy of CFM algorithm. This means that we showed the best performance of the CFM algorithm at the expense of its speed. By observing the following table 2, 3, 4, we can find that the count of edges with backbone degree more than 0.2 is far lower than the count of total edges. If we control threshold f, we can get a fast running speed, but it may be lost a part of the accuracy of CFM algorithm. If users just want to find core communities, they can adjust parameters w and f according their needs.

6.2. Experimental environment

In this section we introduce data sets firstly, then show hardware and software environment of CFM algorithm.

6.2.1. Data Set

We use some standard data set and LFR benchmark(Lancichinetti & Fortunato, 2009): Zachary karate club, American College Football. American College Football and Karate Club are standard data sets, they prove the validity of community detection algorithm. The detailed description of these data sets is shown in table 2, table 3 and table 4.

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.

Netscience-coauthor network data set contains a coauthorship network of scientists working on network theory and experiment, as compiled by M.

Newman in May 2006(Newman, 2006). This network contains total of 1589 scientists and 2742 coauthorships.

Condensed matter collaborations 2003 network contains 31163 vertices and 120029 edges. It is the network of coauthorships between scientists posting preprints on the Condensed Matter E-Print Archive. This version includes all preprints posted between Jan 1, 1995 and June 30, 2003(Newman, 2001). The largest component of this network, which contains 27519 scientists, has been used by several authors as a test-bed for community-finding algorithms for large networks.

LFR benchmark is an implementation of the algorithm described in the paper "Directed, weighted and overlapping benchmark graphs for community detection algorithms", written by Andrea Lancichinetti and Santo Fortunato. In particular, this program is to produce binary networks with overlapping vertices. The order is /benchmark -N n -k 15 -maxk 50 -mu 0.2 -minc 20 -maxc 50 -on 0.1*n -om 4, n is the number of nodes. average degree is 15, maximum degree is 50, mixing parameter is 0.2, minimum for the community sizes is 20, maximum for the community sizes is 50. number of overlapping nodes is 0.1*n. number of memberships of the overlapping vertices is 4. We consider that data set generated by this order can represent a kind of real social network, although it can not represent all kinds of real social network.

6.2.2. Hardware and Software Environment

The server configuration of experimental environment: Dell PowerEdge R720, CPU: Xeon E5-2603*2, memory: 384G(DDR3 1600 16G*24), Hard disk: 300G*2.

Operation System: Windows server 2008, Ubuntu12.0 64bit.

Software: CFinder 2.0.6, SVINET 0.9-beta, CFM 0.1-beta, Louvain method.

CFM algorithm runs on Windows server 2008 and Jdk1.7, it is a Java program. SVINET 0.9-beta runs on Ubuntu12.0 64bit, it is a C++ program. CFinder 2.0.6 runs on Ubuntu12.0 64bit too, it is a Java program. Louvain method runs on Ubuntu14.04 64bit, it is a C++ program.

6.3. Zachary Karate Club Test

We apply CFM algorithm to Zachary karate club network, Figure 1 and figure 2 are the results of applying CFM algorithm based on two expansion measures that defined in definition 5 and 7. CFM algorithm divides this network in two communities in figure 1, and in three communities in figure 2. The division result showed in figure 1 is same with the standard result

Table 2: Data Set description.

Data Set	Vertices	Edges	Known Communities
Zachary's Karate Club	34	78	2
American College Football	115	613	12
Netscience-coauthor	1,589	2,742	Unknown
Condensed matter collaborations	31,163	120,029	Unknown
LFR 1,000	1,000	15,178	24
LFR 2,000	2,000	20,284	52
LFR 5,000	5,000	75,838	129
LFR 10,000	10,000	155,246	264
LFR 20,000	20,000	304,926	521
LFR 50,000	50,000	765,328	1,317
LFR 100,000	100,000	1,529,586	2,668
LFR 20,0000	20,0000	3,057,258	5,275
LFR 50,0000	50,0000	7,657,622	13,127
LFR 100,0000	100,0000	15,315,110	26,374

Table 3: The edges with maximum Backbone Degree.

Data Set	Maximum Backbone Edge	Backbone Degree
American College Football	763,689	1.282727
Zachary's Karate Club	33,31	1.01346
Netscience-coauthor	1430,1429	1.186
Condensed matter collaborations	1886,1885	1.293
LFR 1,000	986, 989	0.802
LFR 2,000	1, 1147	1.123
LFR 5,000	4921,4935	1.123
LFR 10,000	9929,9961	1.281
LFR 20,000	19934,19978	1.264
LFR 50,000	49742,49869	1.264
LFR 100,000	99768,99149	1.272
LFR 20,0000	199811,199658	1.31
LFR 50,0000	499079,497340	1.319
LFR 100,0000	993604,992689	1.057

Table 4: The count of edges with $f \geq 0.2$ and $f \geq 0.3$.

Data Set	$f \geq 0.2$	$f \geq 0.3$
American College Football	449	411
Zachary's Karate Club	31	9
Netscience-coauthor	1196	778
Condensed matter collaborations	1235	244
LFR 1,000	131	46
LFR 2,000	567	227
LFR 5,000	5301	2136
LFR 10,000	11853	4868
LFR 20,000	21,088	8,396
LFR 50,000	55,734	22,497
LFR 100,000	111,992	45,377
LFR 20,0000	218,743	88,131
LFR 50,0000	550,210	223,165
LFR 100,0000	100,945	35,126

of zachary karate club network. We mark these overlapping vertices with multi-colour based on their probability membership in figure 1 and figure 2. Figure 2 shows the more detailed community division than figure 1, and there are less overlapping vertices in figure 2 than in figure 1. We find that CFM can detect more little community base on the expansion defined in definition 7, for example the community 2 is a little community with 3 vertices.

Figure 3 and figure 4 are the results of applying CPM algorithm to zachary karate club network when k=3 and k=4, and different colors represent different communities. When k=3, CPM found 3 communities that including 31 vertices in total. When k=4, CPM found 3 communities that including 12 vertices in total. When k=5, CPM found 1 community that including 6 vertices in total. We do not show these figures that when k is more than 4, because those results can not include all vertices in zachary karate club network. Figure 3 dose not show the clear community division, and the result is different with the standard result. we find that CPM can not give clear community division, but can detect the cliques. We do not give the result of applying MMSB algorithm to zachary karate club network, because MMSB algorithm can not calculate the result of zachary karate club network in more than 3 days, it is a very strange phenomenon, we consider that MMSB has

defects in this data set. Figure 5 is the result of applying louvain method to Zachary Karate Club network. Louvain method divides this network in four communities, two big communities and two small communities, two big communities are similar with the result of CFM algorithm based EX(the short form of definition 5). Two small communities are similar with CFM algorithm based on EXBD(the short form of definition 7) and CPM algorithm when k=3. We find that all algorithms can detect the core cliques of Zachary Karate Club network, but CFM and Louvain method can detect the boundary of Karate Club network more than CPM algorithm, and CFM algorithm can give more accurate division result than CPM and Louvain method in this network.

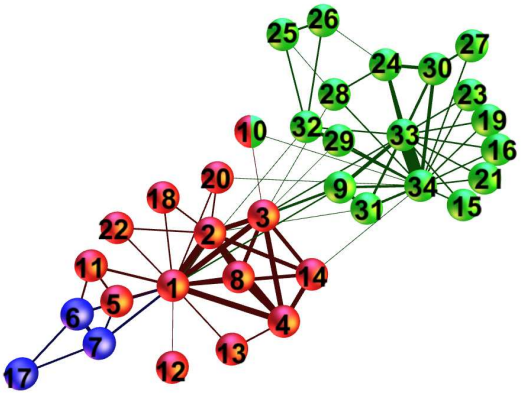

Figure 2: The result of applying CFM algorithm to Zachary Karate Club network base on EXBD, community 0 is marked in green, community 1 is marked in red, community 2 is marked in blue. These multi colored vertices are overlapping vertices.

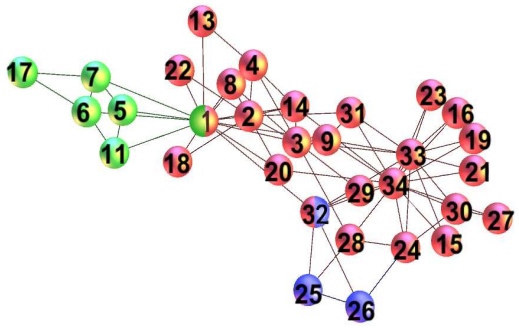

Figure 3: The result of applying CPM algorithm to Zachary Karate Club network when k=3. These multi colored vertices are overlapping vertices.

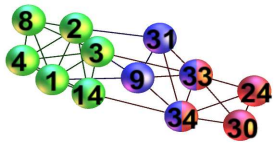

Figure 4: The result of applying CPM algorithm to Zachary Karate Club network when k=4. These multi colored vertices are overlapping vertices.

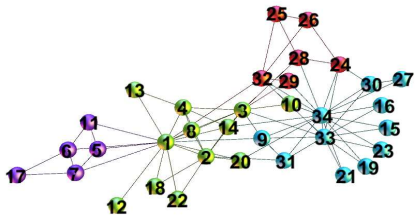

Figure 5: The result of applying Louvain algorithm to Zachary Karate Club network. There are no overlapping vertices in this result.

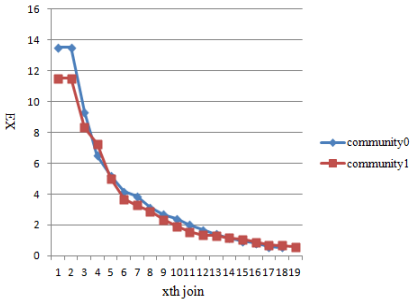

Figure 6: The curve of EX.

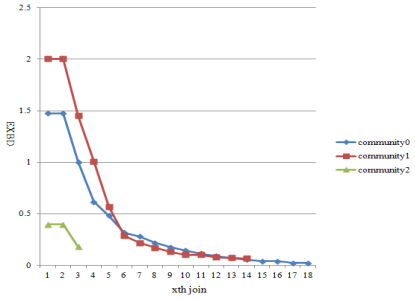

Figure 7: The curve of EXBD.

We display CFM algorithm implementation process to Zachary karate club in table 5 and table 6, table 5 is the division process based on the EX, table 6 is the division process based on EXBD. We find that community 0 and 1 are overlapping on vertex 9, 31, 10 in table 5, and community 0, 1 and 2 are only overlapping on vertex 10 in table 6. We also find that the EX and EXBD of community decreases gradually with the vertex joining, the curves are shown in figure 6 and figure 7, and the vertices's order of joining community is based on sum of backbone degree to the edges of vertex v point to community C that defined in definition 2. This phenomenon is fully verified the definition of overlapping community in this paper. The definition of disjoint community is proved by experiments in our last paper(Yunfeng Xu, 2015), but we only given the definition of community in that paper. Tracking the vertices's order , EX and EXBD to see table 5 and table 6.

Table 5: CFM algorithm implementation process to Zachary Karate Club based on EX.

Vertex ID	Current EX	Community ID	Joining order
34	13.5	0	1
33	13.5	0	1
9	9.333	0	2
31	6.5	0	3
30	5.2	0	4
24	4.166	0	5
32	3.857	0	6
27	3.125	0	7
29	2.666	0	8
28	2.4	0	9
19	2	0	10
23	1.666	0	11
21	1.384	0	12
15	1.143	0	13
16	0.933	0	14
25	0.8125	0	15
26	0.588	0	16
10	0.555	0	17
2	11.5	1	1
1	11.5	1	1
4	8.333	1	2
3	7.25	1	3
8	5	1	4
14	3.666	1	5
9	3.286	1	6
31	2.875	1	7
13	2.333	1	8
22	1.9	1	9
18	1.545	1	10
20	1.333	1	11
5	1.307	1	12
11	1.143	1	13
7	1.067	1	14
6	0.875	1	15
17	0.706	1	16
10	0.667	1	17
12	0.579	1	18

Table 6: CFM algorithm implementation process to Zachary Karate Club based on EXBD.

Vertex ID	Current EXBD	Community ID	Joining order
34	1.475	0	1
33	1.475	0	1
9	1.002	0	2
31	0.618	0	3
30	0.483	0	4
24	0.316	0	5
32	0.281	0	6
27	0.220	0	7
29	0.177	0	8
28	0.144	0	9
19	0.116	0	10
23	0.092	0	11
21	0.073	0	12
15	0.056	0	13
16	0.041	0	14
25	0.040	0	15
26	0.023	0	16
10	0.022	0	17
2	2.005	1	1
1	2.005	1	1
4	1.451	1	2
3	1.009	1	3
8	0.569	1	4
14	0.287	1	5
13	0.220	1	6
22	0.171	1	7
18	0.133	1	8
20	0.104	1	9
11	0.102	1	10
5	0.078	1	11
10	0.072	1	12
12	0.066	1	13
7	0.398	2	1
6	0.398	2	1
17	0.180	2	2

6.4. American College Football Test

We apply CFM, CPM, MMSB and Louvain algorithms to American college football network, Figure 8 and figure 9 are the results of applying CFM algorithm based on two expansion measures that defined in definition 5 and 7. CFM based on EX found 13 communities that including all 115 vertices in American college football network. CFM based on EXBD found 14 communities that including all 115 vertices in American college football network. Figure 10 and figure 11 are the results of applying CPM algorithm to American college football network when k=3 and k=4. When k=3, CPM found 4 communities that including 115 vertices in total. We use bright red, dark red, blue and yellow colors to represent 4 communities. The blue community is completely overlapped by dark red community. We use multi-color to represent overlapping vertices in figure 10. When k=4, CPM found 13 communities that including 113 vertices in total. When k=5, CPM found 15 communities that including 106 vertices in total. When k=6, CPM found 11 communities that including 77 vertices in total. When k=7, CPM found 5 communities that including 41 vertices. We do not show these figures that when k is more than 4, because those results can not include all vertices in American college football network. With the increase of k, the number of vertices in the founded community becomes more less, CPM detects the more core level communities. Figure 12 is the result of applying louvain algorithm to American college football network. Louvain found 9 communities that including 115 vertices in American college football network. Figure 13 is the result of applying MMSB algorithm to American college football network. MMSB found 12 communities that including 112 vertices in American college football network.

The community division result showed in figure 8 is very close to the standard result of zachary karate club network. There are 9 overlapping vertices in figure 8, they are 711, 705, 719, 699, 770, 704, 724, 688 and 748 and marked in pink. There are 6 overlapping vertices in figure 9, they are 728, 732, 766, 727, 705 and 711, and marked in red. There are more overlapping vertices in figure 8 than figure 9, and there are boundary overlap, partial overlap and matryoshka doll overlap in figure 8. There is a overlap between community 12 and community 1, it is a matryoshka doll overlap, because community 12 contains all community 1. There is a partial overlap between community 3 and community 10. There are 6 overlapping vertices in figure 8, these vertices are the boundary of at least 4 communities, this is the boundary overlap. Figure 9 shows the more detailed community division

than figure 8, and there are less overlapping vertices in figure 9 than figure 8. We find that CFM algorithm can detect more little community base on EXBD, for example there are three big communities that contained 2 small communities in figure 9, there are clear boundary between these two small communities, and these structures are hierarchical structures. Hierarchical community structure is one kind of matryoshka doll overlap.

There are not clear boundary in figure 10, figure 11 and figure 13, and we can find very less clear community from these three figures. The community division results of MMSB and CPM algorithms are very different to the standard result of American college football network. There are more clear boundary in figure 12 than figure 10, figure 11 and figure13, but there are less internal structure in figure 12 than figure 8 and figure 9. CFM algorithm can give external boundary and internal structure of community, and the more accurate division result than louvain method, MMSB and CPM. Lovain method can give more clear external boundary of community and more accurate division than CPM and MMSB. The result in American college football network is similar whit the result in Zachary karate club network.

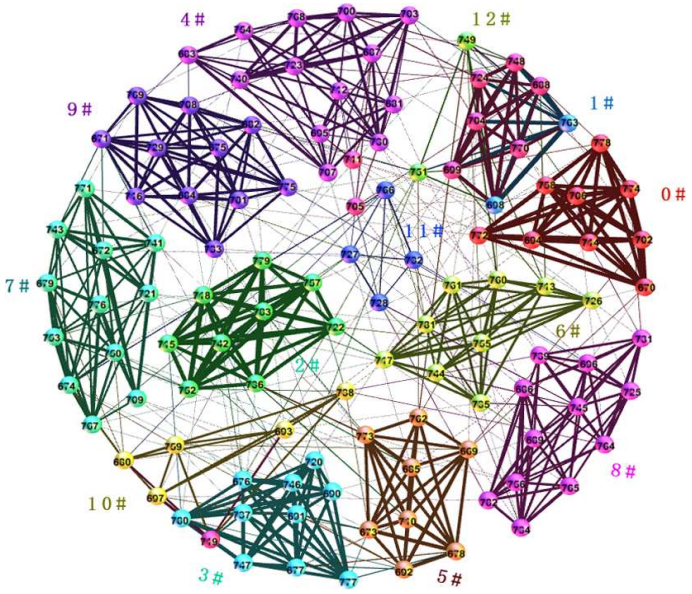

Figure 8: The result of applying CFM algorithm to American College Football network based on EX. These pink vertices are overlapping vertices. They are 711, 705, 719, 699, 770, 704, 724, 688 and 748.

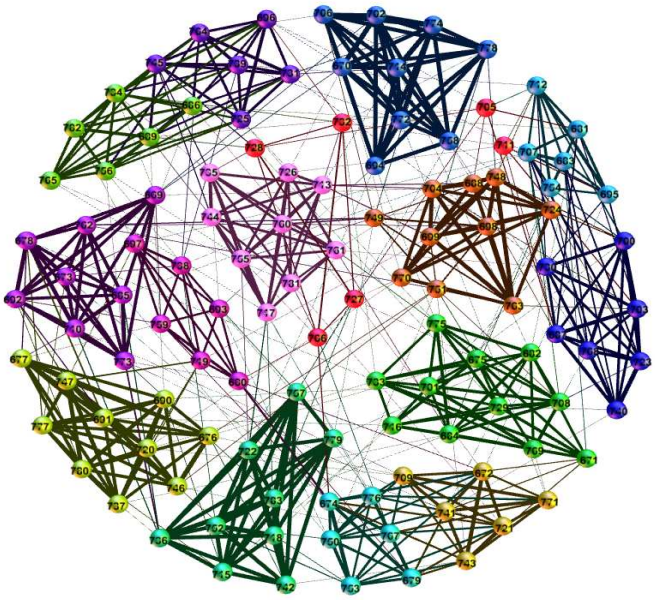

Figure 9: The result of applying CFM algorithm to American College Football network based on EXBD. These red vertices are overlapping vertices.

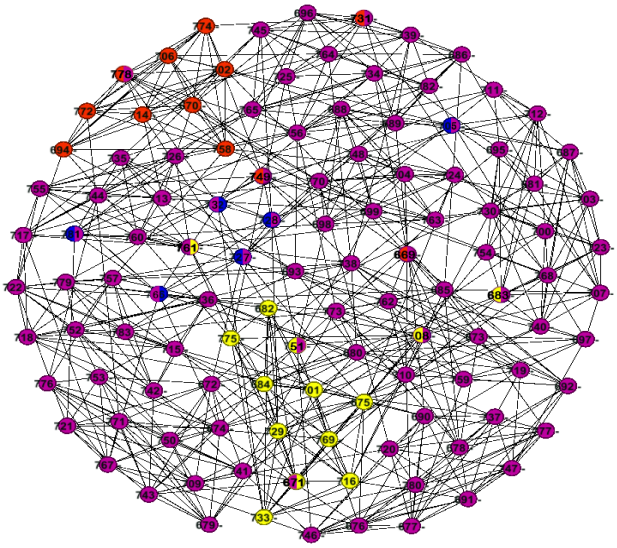

Figure 10: The result of applying CPM algorithm to American College Football network when k=3. These multi colored vertices are overlapping vertices.

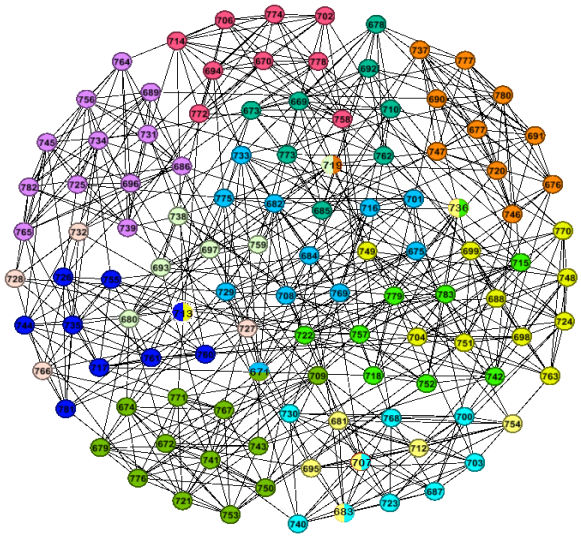

Figure 11: The result of applying CPM algorithm to American College Football network when k=4. These multi colored vertices are overlapping vertices.

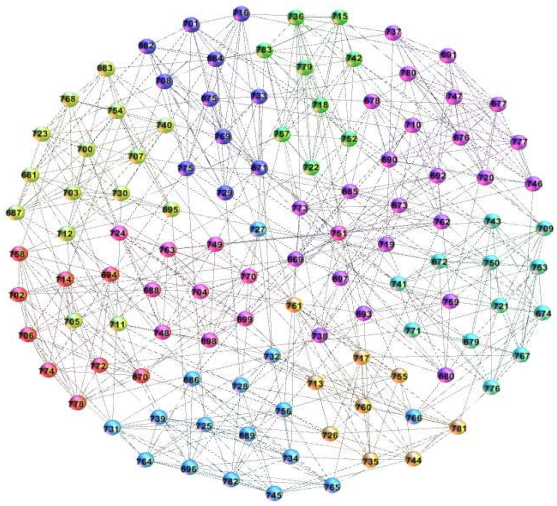

Figure 12: The result of applying louvain algorithm to American College Football network. There are no overlapping vertices in this result.

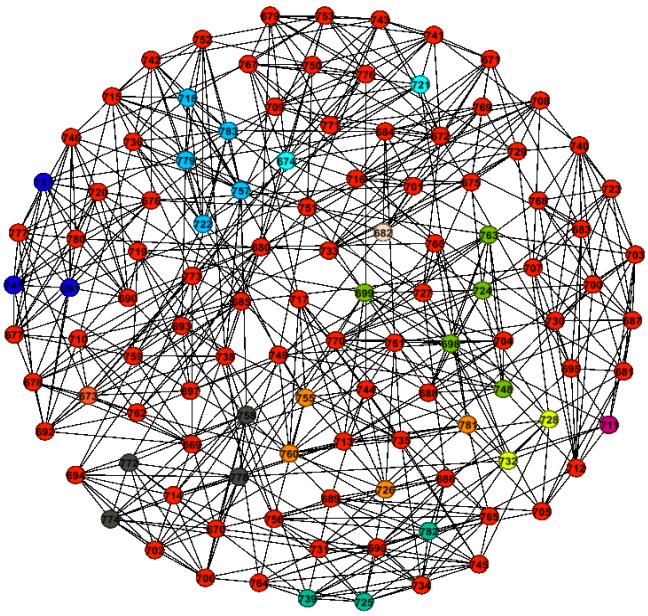

Figure 13: The result of applying MMSB detection algorithm to American College Football network. These red vertices are overlapping vertices.

6.5. Large Real-world Networks Test

We apply CFM, CPM, louvain method and MMSB algorithms to Netscience-coauthor network and Condensed matter collaborations network, and adopt extended modularity Q_m that proposed by Junqiu Li et al.(Li et al., 2014). We cite the experiment results in the work of Junqiu Li et al.(Li et al., 2014) for performance comparison. The Q_m summary of CFM, CPM, Louvain, MMSB and Li's algorithm is shown in table 7. The runtime comparison is not shown, because we have no demo of Li's algorithm, there is no sense to running time comparison on different platform. The Q_m results of CPM $k > 5$ are more lower than CPM $k = 5$, so they are not listed in table 7. The Q_m of CFM is most better than other 4 algorithms, but it is lower than CPM $k = 3$, Louvain, Li's algorithm in American Football. At the same time, we found another problem. The Q_m of CFM in American Football is very low, but we see the partitioning results of CFM in Figure 8 and 9 are very good. Then whether the Q_m can evaluate the quality of community detection results comprehensively, of course, Q_m meets the most needs of community detection evaluation through the work of Junqui et al.(Li et al.,

2014). We will discuss this issue in future work for reasons of space.

Table 7: The Q_m summary of CFM, CPM, louvain method, MMSB and Li's algorithm.

Algorithm	Karate Club	American Football	Netscience	Cond-mat-2003
CFM	0.527	0.305	0.850	0.684
CPM $k = 3$	0.501	0.646	0.700	0.641
CPM $k = 4$	0.062	0.286	0.521	0.354
CPM $k = 5$	0.048	0.204	0.389	0.225
Louvain method	0.431	0.594	0.838	0.457
MMSB	null	0.167	0.256	0.191
Li's algorithm	0.442	0.601	0.842	0.458

6.6. LFR Data Set Test

We apply CFM, CPM, louvain method and MMSB algorithms to LFR data set network. Figure 14 is the precision curve of applying CPM, CFM, louvain method and MMSB algorithms to LFR data set. Figure 15 is the recall curve of applying CPM, CFM, louvain method and MMSB algorithms to LFR data set. Figure 16 is the F-measure curve of applying CPM, CFM, louvain method and MMSB algorithms to LFR data set. We can not get the community division result of MMSB in LFR 200,000, LFR 500,000 and LFR 1000,000, because MMSB ran more than 20 days on PowerEdge R720 in those three LFR data sets, so the curves of MMSB are not complete.

The result of CFM algorithm shows very stable performance in LFR data set from 1000 to 1000,000. Recall, precision and F-measure are all about 60 percent. The results of CPM and MMSB algorithms show jumping performance in recall, precision and F-measure. The precision of CPM algorithm are more high than CFM and MMSB algorithms when $k > 3$, and partly high than CFM and MMSB algorithms when $k = 3$, the curves are shown in figure 14. The recall of CPM algorithm are more high mostly than CFM and MMSB algorithms when $k = 3$, and mostly low than CFM and MMSB algorithms when $k > 3$, the curves are shown in figure 16. Why are the precision and recall curves of CPM like that? The reason is that CPM only detects the core cliques and does not divide all the vertices in network when $k > 3$, this is same with CPM algorithm in Zachary karate club network and American College Football network. The F-measure of CFM algorithm are

more high mostly than CPM, louvain method and MMSB algorithms, only low than CPM algorithm when $k = 3$ and vertex $count > 10,000$, the curves are shown in figure 16. The recall curve of Louvain algorithm is same with CFM algorithm when vertex $count \geq 5,000$, and it is lower than CFM algorithm when vertex $count < 5,000$. The precision curve of louvain method falls rapidly with the increase of vertex count, so the F-measure curve of Louvain algorithm falls rapidly too with the increase of vertex count.

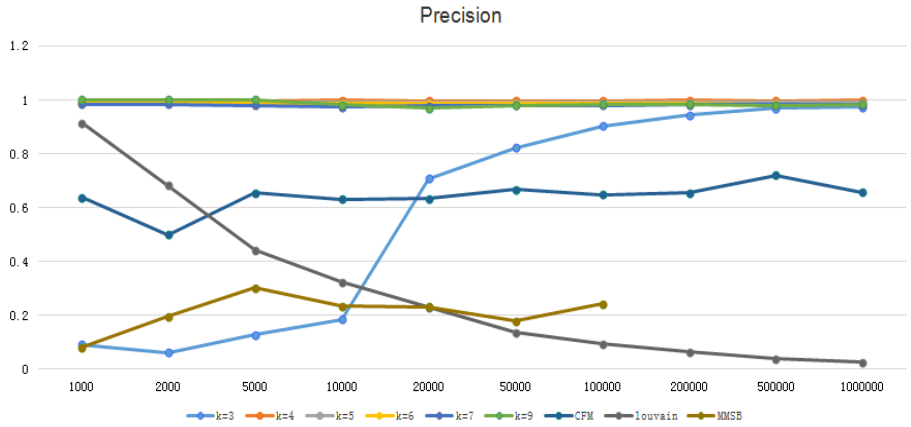

Figure 14: The precision curve of applying CPM, CFM, Louvain and MMSB algorithms to LFR data set.

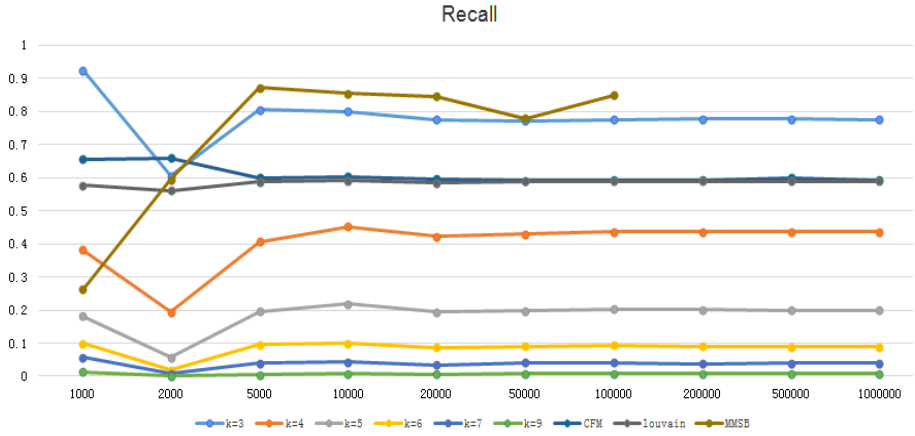

Figure 15: The recall curve of applying CPM, CFM, Louvain and MMSB algorithms to LFR data set.

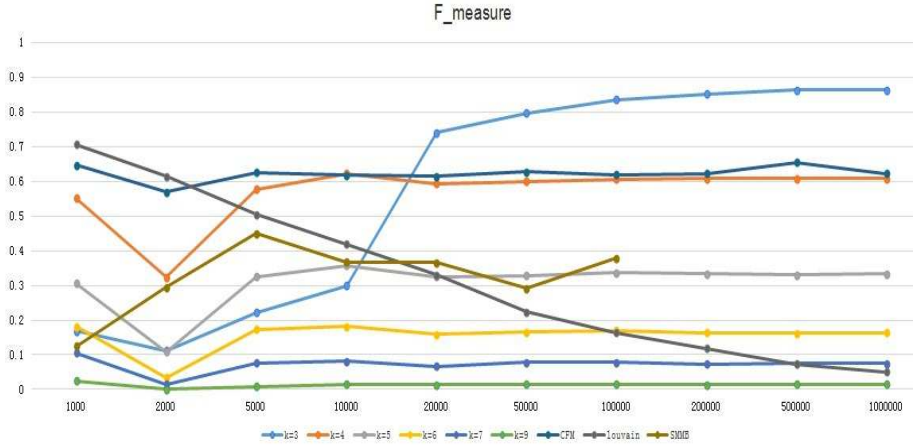

Figure 16: The F-measure curve of applying CPM, CFM, louvain method and MMSB algorithms to LFR data set.

6.7. Run Time Comparition

We compare the runtime of CPM, CFM, louvain method and MMSB on the LFR data set. Table 8 is the runtime list of four algorithm, the time measure is second. Figure 17 is the runtime curve of CPM, CFM, louvain method and MMSB. We find that the runtime curves of CFM and CPM are almost linear growth. A sudden growth arises in the runtime curve of CPM when LFR data set $count > 200,000$, CPM is more than CFM algorithm on LFR500,000 and LFR1000,000. The runtime of louvain method is the lowest in these four algorithms, the runtime of MMSB is the highest. The runtime of CFM algorithm grows linearly with the size of LFR data set. The runtime of MMSB algorithm is not linear growth, and it is more than 20 days on LFR200,000, LFR500,000 and LFR1000,000, so we do not give the complete result of MMSB algorithm, it has lost the meaning of comparison. The runtime of CFM are low than CPM and MMSB when LFR data set $count > 200,000$, so we consider that CFM can deal with large scale data.

Table 8: The runtime list of CPM, CFM, louvain method and MMSB.

LFR Data Set	CPM(s)	CFM(s)	MMSB(s)	Louvain method(s)
LFR 1,000	1	6	12	1
LFR 2,000	1.5	7	16	1
LFR 5,000	2	26	120	2
LFR 10,000	2	61	222	2
LFR 20,000	4	113	1136	3
LFR 50,000	10	266	5861	3
LFR 100,000	21	710	58641	6
LFR 20,0000	27	1,340	null	16
LFR 50,0000	5,646	4,039	null	22
LFR 100,0000	11,326	9,037	null	39

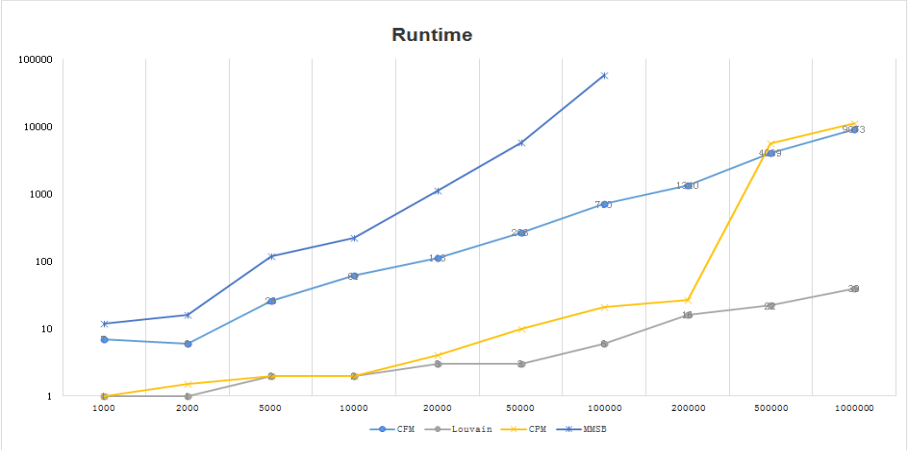

Figure 17: The runtime curve of applying CPM, CFM, MMSB and louvain method to LFR data set, the horizontal axis is the count of LFR data set, the vertical axis is the $log_{10}Runtime$.

6.8. Discussion

Through above experiments, we found that CFM algorithm has better accuracy and effectiveness than CPM, louvain method and MMSB in zachary karate club and American college football data sets, and it can get good score of F-measure and runtime. And CFM can get slightly better performance on the metric Q_m than other four algorithms in large real world social networks,

but it is low in the American Football data set. The recall, precision , F-measure and runtime of CPM algorithm are all linear growth with the growth of LFR data set count, this is a perfect performance, this feature is similar to CPM, but it is better than MMSB. MMSB adopts variational inference algorithm to approximate the posterior inference of the mixed membership, it need many iterations, the computing cost of it is very high, so the runtime is more than CPM, Louvain and CFM. CFM adopted backbone degree, EX and EXBD to measure the social network, so the result is only one. CPM adopted k-clique to measure the social network, then got many results, for example k=3,4,..., these results are discrete. CFM algorithm based on EXBD can give more detailed community division than CFM based on EX in zachary karate club and American college football data sets, and CFM algorithm based on EXBD can find matryoshka doll overlap(hierarchical community structures), boundary overlap and partial overlap in real-world social networks. This proved that EXBD metric can improve the resolution of community detection. The runtime of louvain method is the lowest in these four algorithms, the runtime of MMSB is the highest. The precision and F-measure of louvain method decreases linearly with the growth of LFR data set count, but louvain method can get good performance in zachary karate club and American college football data sets, so we consider that louvain method is better on dealing with real social networks than dealing with other networks.

Though the experiments to LFR data set, we find that CFM algorithm is superior than MMSB and louvain method to discover the structure of social networks, and the demand for memory and runtime is less than the MMSB algorithm, and CFM algorithm is mostly superior than CPM algorithm on F-measure when $k > 3$, partly lower than CPM algorithm when $k = 3$. The runtime of CFM algorithm is lower than CPM algorithm when LFR data set $count \geq 500,000$, so CFM can deal with more larger scale social network than CPM. The F-measure curve of louvain method falls rapidly too with the increase of vertex count. In summary, CFM algorithm has better performance than CPM, louvain method and MMSB in large scale social network.

LFR data sets are not real social networks, they are generated by a fixed order with the same parameters, and they become more and more sparse with the increase of the network size, so the precision of CPM grows rapidly with the growth of LFR data set count when k=3, and we find that the communities detected by CPM become more and more smaller through the observation to the division results of LFR data sets, it means that there are

no 3-cliques between these communities. CPM only found the communities that included core cliques, it do nothing to the boundary vertices that not belonged any 3-cliques. So although CPM has achieved good results in F-measure when k=3, but we don't think the CPM is better than CFM on LFR data sets. And Yang et al. found that when k=4, CPM can get more good result than k equals other values (Yang & Leskovec, 2013), but the F-measure of CPM when k=3 is partly higher than when k=4. So we consider that F-measure are not a perfect metric to overlapping community detection, but it can be a reference.

Summing up the above four kinds of evaluation results, we can have a more comprehensive understanding of these five algorithms. We divide the evaluation results of the four algorithms into 3 levels: low, medium and high, the result is shown in table 9. These 3 levels are based on the relative comparison of the above three kinds of evaluation results, the detail is shown in above figures and tables. The F-measure of CPM k=3 is medium, because it is partly better than CFM and partly lower than CFM. We think CFM is the better than other three algorithms based on the comprehensive performance. Li's algorithm is not complete in table 9, because we have no demo of Li's algorithm, so that there is no complete experimental results about it.

Table 9: The evaluation summary of CFM, CPM, louvain method, MMSB and Li's algorithm.

Algorithm	Effectiveness and Accuracy	F-measure	Runtime	Q_m
CFM	High	High	Medium	High
CPM $k = 3$	Low	Medium	Medium	High
CPM $k = 4$	Medium	High	Medium	Medium
CPM $k > 4$	Low	Low	Medium	Low
Louvain	Medium	Medium	High	High
MMSB	Low	Medium	Low	Low
Li's algorithm	null	null	null	High

7. Conclusions

In this paper, we focus on the problem of overlapping community detection in social networks which is the key tool for understanding the function of the networks and its structure. Many algorithms in this area are developed

and we have discussed their limits in this paper. Our main contributions are three folders. Firstly we extend the community forest model to overlapping community forest model and disjoint community forest model based on these social and biological properties. Secondly we mainly propose the definition of overlapping community and disjoint community based on EX, EXBD and backbone degree. Thirdly we develop CFM algorithm based on EX or EXBD to discover overlapping communities from real social networks.

CFM algorithm can deal with the four challenges to existing works. Firstly we give the definition of community with clear boundaries and internal structure in Model and Problem Formalization section of this paper. And through experiments we proved that CFM algorithm based on this definition has good performance for overlapping community detection. Secondly there are no complicated optimizations of the likelihood function in CFM algorithm. More detailed algorithm implementation is shown in Algorithm section in this paper. Thirdly there are no iterative and sampling in CFM algorithm. We have implemented CFM algorithm based on hadoop platform, but this paper focuses on algorithms and models, and we will introduce parallel CFM algorithm in subsequent articles. Finally CFM algorithm mines the social and biological characteristics of social network in-depth, and calculate simple and effective measures such as expansion and backbone degree, the process of community detection does not need complicated physical and mathematical methods. These four points are the main difference between our work and existing works.

CFM algorithm is different with all community detection algorithms that mentioned in this paper, because it is based on a biological and sociological model named community forest model, and CFM algorithm is a simple and direct approach to detect community in networks, it integrated EX/EXBD and backbone degree. CFM algorithm has better performance than MMSB algorithm, louvain method, CPM algorithm and Li's algorithm in large scale social network. It works well on American college football, karate club, Netscience-coauthor, Condensed matter collaborations, LFR etc. data sets.

In addition to the above contributions, our work can also be promoted as follows. Firstly, the definitions of overlapping community forest model and disjoint forest model are based on EX, EXBD and probability membership, so they are all defined from the perspective of community and vertex. We do not discuss the detailed overlapping degree of communities in our models for space limit, we will give more perfect definition to this question in next paper. Secondly, we used Q_m and F-measure to evaluate these five algorithms.

Q_m is a reasonable metric to large real world social networks that without standard partition results. But we think that it has room for improvement through experiments. Improved F-measure in this paper is not the perfect metric to evaluate the overlapping community detection results, it can be a reference. Although Amig et al. proved that only BCubed(Bagga & Baldwin, 1998) satisfies all formal constraints. But the problem of overlapping community detection has its own unique characteristics, for example these social and biological characteristics that pointed out in this paper. We will integrate BCubed metric(Amig et al., 2009), backbone degree, network community profile(Leskovec et al., 2010), Q_m etc. for evaluation in our future work. Thirdly, we used some real world networks and LFR data set. LFR data set is not real social network, but there are no large scale social networks that with standard community detection results, so LFR data set is a compromise choice. We will adopt more LFR data sets with more various parameters to evaluate the performance of these four algorithms in the future work, not only LFR data set with the one set of parameters in this paper. Because more various parameters can generate more similar network with real social network. Finally, CFM algorithm has good performance to social networks, but it only detects communities in a single-machine environment in undirected networks currently , and there are some little defects such as that runtime is greater than louvain method.

Our next work is to optimize CFM algorithm for the above mentioned problems, and adjust it to suit for detecting overlapping communities in more large-scale networks.

8. Acknowledgments

This work is supported by National Basic Research Program of China(973 Program)(Grant No2012CB316301). And we also got the support from National Natural Science Foundation of China(Grant No:71271076, 61300120, 51271033, 61272362), Hebei science and technology support program(Grant No:16210312D) and Hebei University Of Science and Technology enterprise docking project(Grant No:2016JSDJ11).

References

Ahn, Y., Bagrow, J., & Lehmann, S. (). Communities and hierarchical organization of links in complex networks. 2009. *Available from: arxiv*, *903*.

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).

Amig, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, *12*, 461–486.

Artiles, J., Gonzalo, J., & Sekine, S. (2007). The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *In SemEval 2007, ACL* (pp. 64–69).

Bagga, A., & Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1* (pp. 79–85).

Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., & Mooney, R. J. (2005). Model-based overlapping clustering. *In Proc. of 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD-05*, (pp. 532–537).

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory & Experiment*, *30*, 155–168.

Burt, R. S. (1994). Structural holes: The social structure of competition. *Economic Journal*, .

Connell, J. H., & Slatyer, R. O. (1977). Mechanisms of succession in natural communities and their role in community stability and organization. *American Naturalist*, *111*, 1119–1144.

Cui, Y., & Wang, X. (2014). Uncovering overlapping community structures by the key bi-community and intimate degree in bipartite networks. *Physica A Statistical Mechanics & Its Applications*, *407*, 7–14.

Cui, Y., Wang, X., Cui, Y., & Wang, X. (2016). Detecting one-mode communities in bipartite networks by bipartite clustering triangular. *Physica A-Statistical Mechanics and Its Applications*, *457*, 307–315.

Cui, Y., Wang, X., & Eustace, J. (2014a). Detecting community structure via the maximal sub-graphs and belonging degrees in complex networks. *Physica A Statistical Mechanics & Its Applications*, *416*, 198–207.

Cui, Y., Wang, X., & Li, J. (2014b). Detecting overlapping communities in networks using the maximal sub-graph and the clustering coefficient. *Physica A Statistical Mechanics & Its Applications*, *405*, 85–91.

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Eustace, J., Wang, X., & Cui, Y. (2015a). Community detection using local neighborhood in complex networks. *Physica A Statistical Mechanics & Its Applications*, *436*, 665–677.

Eustace, J., Wang, X., & Cui, Y. (2015b). Overlapping community detection using neighborhood ratio matrix. *Physica A Statistical Mechanics & Its Applications*, *421*, 510–521.

Eustace, J., Wang, X., & Li, J. (2014). Approximating web communities using subspace decomposition. *Knowledge-Based Systems*, *70*, 118–127.

Evans, T., & Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Physical Review E*, *80*, 016105.

Evans, T. S. (2010). Clique graphs and overlapping communities. *Journal of Statistical Mechanics Theory & Experiment*, *2010*, 257–265.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*, 75–174.

Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Science*, *104*, 36–41.

Gan, W. Y., Nan, H. E., De-Yi, L. I., & Wang, J. M. (2009). Community discovery method in networks based on topological potential. *Journal of Software*, *20*, 2241–2254.

Gopalan, P. K., & Blei, D. M. (2013). Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 14534–14539.

Han, Y., Li, D., & Wang, T. (2011). Identifying different community members in complex networks based on topology potential. *Frontiers of Computer Science in China*, *5*, 87–99.

Havemann, F., Heinz, M., Struck, A., & Gl?ser, J. (2010). Identification of overlapping communities and their hierarchy by locally calculating community-changing resolution levels. *Computer Science*, *2011*, 01023.

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

Kelley, & Stephen (2009). The existence and discovery of overlapping communities in large-scale networks. *Dissertations & Theses - Gradworks*, .

Kelley, S., Goldberg, M., Magdon-Ismail, M., Mertsalov, K., & Wallace, A. (2012). *Defining and Discovering Communities in Social Networks*. Springer US.

Kim, Y., & Jeong, H. (2011). The map equation for link community. *Corr*, *84*, 1402–1409.

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

Lancichinetti, A., & Fortunato, S. (2009). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, *80*, 016118.

Lancichinetti, A., Fortunato, S., & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, *11*, 033015.

Latouche, P., Birmelé, E., & Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, (pp. 309–336).

Lei, X., Wu, S., Ge, L., & Zhang, A. (2013). Clustering and overlapping modules detection in ppi network based on ibfo. *Proteomics*, *13*, 278–90.

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.

Li, J., Wang, X., & Cui, Y. (2014). Uncovering the overlapping community structure of complex networks by maximal cliques. *Physica A Statistical Mechanics & Its Applications*, *415*, 398–406.

Li, J., Wang, X., & Eustace, J. (2013a). Detecting overlapping communities by seed community in weighted complex networks. *Physica A Statistical Mechanics & Its Applications*, *392*, 6125–6134.

Li, Y., Wang, H., Li, J., & Gao, H. (2013b). Efficient community detection with additive constrains on large networks. *Knowledge-Based Systems*, *52*, 268–278.

Lou, T., & Tang, J. (2013). Mining structural hole spanners through information diffusion in social networks. In *International Conference on World Wide Web* (pp. 825–836).

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

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, *74*, 92–100.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*, 026113.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Working Papers*, *98*, 404–9.

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.

Reid, F., McDaid, A., & Hurley, N. (2011). Partitioning breaks communities. *Lecture Notes in Social Networks*, (pp. 102 – 109).

Ren, G., & Wang, X. (2014). Epidemic spreading in time-varying community networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *24*, 023116.

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

Shen, H., Cheng, X., Cai, K., & Hu, M.-B. (2009a). Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, *388*, 1706–1712.

Shen, H. W., Cheng, X. Q., & Guo, J. F. (2009b). Quantifying and identifying the overlapping community structure in networks. *Journal of Statistical Mechanics Theory & Experiment*, *2009*, 07042.

Wang, X., & Li, J. (2013). Detecting communities by the core-vertex and intimate degree in complex networks. *Physica A Statistical Mechanics & Its Applications*, *392*, 2555–2563.

Wang, Z. X., Li, Z. C., Ding, X. F., & Tang, J. H. (2016). Overlapping community detection based on node location analysis. *Knowledge-Based Systems*, *105*, 225–235.

Whang, J. J., Gleich, D. F., & Dhillon, I. S. (2015). Overlapping community detection using neighborhood-inflated seed expansion. *IEEE Transactions on Knowledge & Data Engineering*, *28*, 1272–1284.

Whang, J. J., Gleich, D. F., & Dhillon, I. S. (2016). Overlapping community detection using neighborhood-inflated seed expansion. *IEEE Transactions on Knowledge & Data Engineering*, *28*, 1272–1284.

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. (2013). Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 587–596). ACM.

Yunfeng Xu, H. X., Dongwen Zhang (2015). A novel disjoint community detection algorithm for social networks based on backbone degree and expansion. *Expert Systems with Applications*, *42*, 8349–8360.

Z, D., X, Z., D, S., & B, L. (2016). Overlapping community detection based on network decomposition. *Scientific reports*, *6*.

Zhao, Z., Feng, S., Wang, Q., Huang, J. Z., Williams, G. J., & Fan, J. (2012). Topic oriented community detection through social objects and link analysis in social networks. *Knowledge-Based Systems*, *26*, 164–173.