# Springer Theses

## Recognizing Outstanding Ph.D. Research

For further volumes:
www.springer.com/series/8790

## Aims and Scope

The series "Springer Theses" brings together a selection of the very best Ph.D. theses from around the world and across the physical sciences. Nominated and endorsed by two recognized specialists, each published volume has been selected for its scientific excellence and the high impact of its contents for the pertinent field of research. For greater accessibility to non-specialists, the published versions include an extended introduction, as well as a foreword by the student's supervisor explaining the special relevance of the work for the field. As a whole, the series will provide a valuable resource both for newcomers to the research fields described, and for other scientists seeking detailed background information on special questions. Finally, it provides an accredited documentation of the valuable contributions made by today's younger generation of scientists.

## Theses are accepted into the series by invited nomination only and must fulfill all of the following criteria

- They must be written in good English.
- The topic should fall within the confines of Chemistry, Physics, Earth Sciences, Engineering and related interdisciplinary fields such as Materials, Nanoscience, Chemical Engineering, Complex Systems and Biophysics.
- The work reported in the thesis must represent a significant scientific advance.
- If the thesis includes previously published material, permission to reproduce this must be gained from the respective copyright holder.
- They must have been examined and passed during the 12 months prior to nomination.
- Each thesis should include a foreword by the supervisor outlining the significance of its content.
- The theses should have a clearly defined structure including an introduction accessible to scientists not expert in that particular field.

Hua-Wei Shen

# Community Structure of Complex Networks

Springer

*Author*
Dr. Hua-Wei Shen
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China

*Supervisor*
Prof. Xue-Qi Cheng
Institute of Computing Technology
Chinese Academy of Science
Beijing, China

ISSN 2190-5053 ISSN 2190-5061 (electronic)
Springer Theses
ISBN 978-3-642-31820-7 ISBN 978-3-642-31821-4 (eBook)
DOI 10.1007/978-3-642-31821-4
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012956465

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*To my dearest wife Yü-Yü Cheng, and our lovely son Jun-Hao Shen.*

# Supervisor's Foreword

We are now surrounded by the ubiquitous complex networks, varying from the World Wide Web to the booming online social networks, from the power grid to the transportation network, from the communication networks to various economic networks, from the networks within cell or organism to the network of ecosystem. These networks are from various disciplines, constructed for different purposes, and seem to be unrelated to each other. However, these networks exhibit astounding general characteristics, including the small-world phenomenon and the power-law degree distribution. It is the right time to say that "the network takeover". Meanwhile, comprehensive research on complex networks requires the revolution of methodology. The reductionism, as a paradigm, is expired. The graph theory, as a mathematical discipline, hits the limit. From the perspective of complexity and with the network thinking, data-driven methodology is developing into a new discipline, i.e., network science.

As a salient structural characteristic of complex networks, community structure indicates the regularity of topological structure and reflects the locality of relationships among the components of networks. Community structure is fundamental to many functional features of complex networks, such as the robustness and navigability. Moreover, community structure affects or even determines the behavior of the dynamical processes taking place on networks, including the information diffusion, the spread of disease and rumor, and synchronization. Therefore, community structure is crucial to understanding the relation between the structure and function of complex networks and has important theoretical and practical implications to utilize and control the dynamics on, or of, the complex networks.

This book focuses on the community structure of complex networks. In particular, this book provides a clear review for the research advances in the community detection of networks, which is one of the hottest research topics in network science. In this book, the author studies four critical aspects of community structure. The four aspects are the overlaps among communities, the multiscale of community structure, the relationship between community structure and network dynamics, and the coexistence of multiple types of structural regularities beyond community structure. Aiming to investigate the community structure in real world networks, this book first highlights the limitation of modularity optimization, which is a classic method for community detection. The author first proposes the algorithm to simultaneously detect the overlapping and hierarchical community structure. Then, the author proposes a dimensionality reduction framework for uncovering the multiscale community structure in networks with heterogeneous networks. Further, the author studies the diffusion dynamics on networks and reveals the relationship between the table transients in diffusion process and the intrinsic community structure in networks. Finally, the author explores the multiple types of structural regularities in networks using probabilistic graphical model. Most of the preliminary works of this book have been published in prestigious journals on network science, e.g., the Physical Review E, and Journal of Statistical Mechanics. This book is also heavily based on Dr. Shen's doctoral thesis, but with a substantial expansion based on his follow-up research. Dr. Shen's thesis was completed in Research Center for Research Center for Web Data Science and Engineering, Institute of Computing Technology, Chinese Academy of Sciences. This thesis was honored with the "Top 100 Excellent Doctoral Dissertations Award" from the Chinese Academy of Sciences and was nominated as the "Outstanding Doctoral Dissertation" by the Chinese Computer Federation. In general, this book brings together the recent research efforts of Dr. Shen in the field of community structure in networks.

I believe both the researchers and practitioners in the field of social network analysis and the broader area of network science can benefit from reading this book. Moreover, this book shows the research track of Dr. Shen from a Ph.D. student to a professor, which may be of interest particularly to new Ph.D. students. I want to compliment Dr. Shen for having written such an outstanding book for the network science community.

Institute of Computing Technology, Beijing, China Xue-Qi Cheng

# Acknowledgements

This book was partially supported by the National Natural Science Foundation of China (NSFC) under grant Nos. 61232010, 61202215, 60873245, 60872036, 60933005, 60873217, and 60803123, and the National Basic Research Program of China (the 973 program) under grant Nos. 2004CB318109, 2007CB310805, and 2012CB316303, and the National High-Tech R&D Program of China (the 863 program) under grant Nos. 2006AA01Z452 and 2010AA012502, and the Beijing Natural Science Foundation under grant No. 4122077. This book was also supported by Key Lab of Information Network Security, Ministry of Public Security.

The author would also like to thank the supports from various research centers, including: (1) Research Center of Web Data Science and Engineering, Institute of Computing Technology, Chinese Academy of Sciences; (2) Key Lab of Information Network Security, Ministry of Public Security.

# Parts of This Book Have Been Published in the Following Articles

Shen, H.W., Cheng, X.Q., Cai, K., Hu, M.B.: Detect overlapping and hierarchical community structure in networks. Physica A **388**(8), 1706–1712 (2009) (Reproduced with Permission)

Shen, H.W., Cheng, X.Q., Guo, J.F.: Quantifying and identifying the overlapping community structure in networks. J. Stat. Mech. P07042 (2009) (Reproduced with Permission)

Shen, H.W., Cheng, X.Q., Fang, B.X.: Covariance, correlation matrix, and the multiscale community structure of networks. Phys. Rev. E **82**, 016114 (2010) (Reproduced with Permission)

Shen, H.W., Cheng, X.Q.: Spectral methods for the detection of network community structure: a comparative analysis. J. Stat. Mech. P10020 (2010) (Reproduced with Permission)

Shen, H.W., Cheng, X.Q.: Uncovering the community structure associated with the diffusion dynamics on networks. J. Stat. Mech. P04024 (2010) (Reproduced with Permission)

Shen, H.W., Cheng, X.Q., Guo, J.F.: Exploring the structural regularities in networks. Phys. Rev. E **84**, 056111 (2011) (Reproduced with Permission)

Shen, H.W., Cheng, X.Q., Wang, Y.Z., Chen, Y.: A dimensionality reduction framework for detection of multiscale structure in heterogeneous networks. J. Comput. Sci. Technol. **27**, 341–357 (2012) (Reproduced with Permission)

# Contents

# Chapter 1
# Community Structure: An Introduction

## 1.1 Network Science: An Emerging Discipline

Nature and society are composed of a wide variety of complex systems with very different scales. These systems range from cell to ecosystem, from the Internet to the Web, from power grid to various communication systems, from stock markets to other economic systems. Distinguishing from simple systems where the strength of interaction is uniquely determined by the physical distance, the components of complex systems highly interact with each other in the way unconstrained by certain distance measurements. These interactions influence and even determine the function and behavior of these complex systems. The whole system is not the simple aggregation of all these components. The system itself exhibits collective characteristics which are distinct from individual behavior. The collective behavior is emergent from spontaneous individual behaviors. We can see that disorder and order coexist in complex systems. To understand the function and behavior of complex systems, we need to study the pattern of interactions among components [1].

Network provides a powerful mathematical tool to represent and study complex systems [2]. For example, the scientific literature can be represented as a network of articles connected by citation relationships; the Web is a vast information network of Web pages linked by hyperlinks; the Internet is a network of routers or autonomous systems connected by various physical links or wireless links; society is a complex network where nodes are individuals and links correspond to various social relationships; the cell is depicted as networks of chemicals linked by chemical reactions; the stock market is best described as a network of traders linked by trading relationships. The underlying networks for these complex systems exhibit non-trivial topological characteristics. It requires considerable efforts to understand the structure of these complex networks and to provide some insights for understanding of the function of networks [3].

Network is absolutely not a new concept. Actually, the study of network has a long history and can date back to Euler's solution of the puzzle of Köigsberg's bridges in 1736 [4]. Since then, graph theory is gradually formed and has developed an arsenal of successful tools to study the properties of networks [5]. As the

H.-W. Shen, *Community Structure of Complex Networks*, Springer Theses,
DOI 10.1007/978-3-642-31821-4_1, © Springer-Verlag Berlin Heidelberg 2013

most prominent development with respect to network in the last century, random networks, developed by Erdős and Rényi, place us in an ultimately random universe [6]. Meanwhile, scientists who do not believe in the wholly-random universe begin to investigate real world networks from various fields. Such kind of empirical studies gradually terminates the random universe for network and finally leads to the birth of a new discipline—network science [2, 3, 7–9].

The emergence of network science is contributed to two critical historic opportunities: the unprecedent availability of data from various fields and the increased computing power or computational resources. In addition, the breakdown of boundaries between disciplines enables scientists to share diverse datasets and communicate ideas from different disciplines. This allows them to uncover the generic properties of complex networks.

Network science aims to investigate the universal properties of networks, to study the mechanism underlying the formation and growth of networks, to find the intrinsic laws dominating the universe of networks. Three well-known generic properties of complex networks are respectively power-law degree distribution [10], small world [11], and the high clustering coefficient [11]. In this monograph, we will focus on the community structure of network, which is another salient and common structural characteristic of complex network [12, 13].

## 1.2 Community Structure: An Salient Structural Characteristic of Networks

We know that real networks are not random and they usually exhibit inhomogeneity, indicating the coexistence of order and organization. For example, the power-law degree distribution characterizes the inhomogeneity of node degrees, i.e., a few nodes with very high degree coexist with many nodes with low degree. Furthermore, the distribution of links also shows inhomogeneity, globally and locally, describing the phenomenon that nodes naturally cluster into groups and links are more likely to connect nodes within the same group. This phenomenon tells us that the organization of network is modular. Network scientists call this phenomenon as community structure of networks [14].

Community structure embodies the famous saying that “the birds of a feather flock together”. In society, individuals with similar interests are more likely to become friends [15, 16]. In the Web, web pages with related topics are often hyperlinked together [17]. In the protein interaction network, communities are composed of proteins with the same specific function for chemical reactions [18, 19]. In metabolic networks, communities may correspond to functional modules such as cycles and pathways [20]. In food webs, compartments can be viewed as communities [21, 22].

Communities in networks are crucial to understand the organization principle and the structural regularities of networks. For the Web, organizing the web pages with related topics into communities is convenient to Web surfers to efficiently browse the Web. For the Internet, communications are often conducted within the communities which correspond to computers in the same autonomous system. In online shopping sites, customers with similar purchase history are viewed as communities and these communities are often used to improve the effectiveness of recommendation systems. For self-organizing networks, community is critical to efficient decentralized navigation and it is often used to guide the design of routing tables that specify how nodes have to communicate to other nodes. Furthermore, communities are also helpful to network visualization and network compression.

Furthermore, community structure is important since it is closely related to the hierarchical organization of many complex systems in the real world. For example, the organization of company is hierarchical: a company is composed of several departments and each department may comprise in several groups and so on. For human body, the body is composed by organs and organs are composed by tissues. These hierarchical organization is corresponds the hierarchical community structure, i.e., networks are composed by communities including smaller communities, which in turn include smaller communities, etc. Such kind of hierarchical organization provides a way to make the system function efficiently and effectively. In this hierarchical organization, each subpart can be improved by adopting new technology independently. Also, it reduces the possibility that errors or failures can cascade in the whole systems.

Because of the important implication of community structure, the community structure has attracted much academic and industrial attention from various fields. In 2002, Girvan and Newman first investigate the community structure in social and biological networks [14]. In their seminal paper, the communities are identified in a divisive way, where links are deleted iteratively according to the measurement "edge betweenness". This measurement quantifies the importance of the role of the edges in bridging the communication of signals transmitted along the shortest path. Since then, the community structure becomes one of the hottest research topics in network science. The participation of physicists brings about the method of spin models, optimization, percolation, random walks, and synchronization. Meanwhile, the scientists in computer science and machine learning provide us many efficient algorithms and techniques for the identification of network communities. The study on community structure has also taken advantage of concepts and methods from computer science, nonlinear dynamics, sociology, discrete mathematics. In what follows, we will give a brief review for the development on community detection.

## 1.3 A Brief Review for Community Detection

Many methods for community detection have been proposed and successfully applied to several specific networks [12]. Each method has a specific definition to community or has certain understanding or explanation to the implication of community structure in networks. In this section, we do not aim to give a thorough survey for the development of community detection methods. We just want to give the readers a brief introduction to community detection and to help readers to grasp the main aspects of community detection.

### *1.3.1 What Is a Community?*

The most fundamental question for community detection is "what is a community". Different answers to this question will lead to different community detection methods. Unfortunately, community is only a qualitative concept and there is no widely-accepted quantitative definition to community until now. Generally speaking, the definition to community depends heavily on the specific context and the application demand. Moreover, several researchers just take the output results of their community detection methods as community and do not give any definition or description to their obtained communities. In general, communities of network are groups of nodes within which nodes are much more connected to each other than to the rest part of the network. Based on the different perspective of the definition to community, definitions to community can be roughly classified into two categories, namely local definition and global definition. In what follows, we will introduce the three kinds of definitions and several classic definitions.

#### 1.3.1.1 Local Definitions to Community

Local definitions to community define a community only according to the information of the community itself. Specifically, a node group is defined as a community by giving some required properties of the group or by setting some constraints to the group. According to the links considered in the definition of community, local definitions to community can be further classified into two categories.

The first category of local definitions focuses only on the inner links of community. In general, a community is defined as a node group which satisfies certain constraint and which is not the subset of any other group which also satisfies this constraint. According to the link density of a node group, the link pattern with the highest link density is clique. Thus, community can be defined as maximal clique. However, the rigid requirement of link density defies such a definition. To combat this problem, several kinds of relaxation to the definition are proposed. Palla et al. proposed to use clique percolation to define community [23]. This kind of definition to community is a generalization to connected component of networks. Meanwhile, the definition based on clique percolation can avoid the high requirement for link density of maximal cliques. Furthermore, researchers also proposed several other relaxations. The $n$-clique based community [24] requires that the distance between any two nodes in the same community is no more than $n$. Note that the $n$-clique based community cannot guarantee that the diameter of a community is no more than $n$. The reason is that two nodes in the same community may reach each other along the path containing the nodes outside the community. To combat this problem, several variations, e.g., $n$-clan and $n$-club [25] are proposed. Furthermore, the constraint is further relaxed from constraints on any pair of nodes to constraints on the relationship between each node with all the other nodes in the community. Specifical examples are $k$-plex [26] and $k$-core [27], which are widely used in sociometric and social network analysis. Taking the $k$-core as an example, $k$-core based community requires that each node in the community have links to at least $k$ nodes in the same community. In addition, $k$-core is equivalent to $p$-quasi clique [28].

The second category of local definitions considers both the inner links of community and the links between the community and the rest part of networks. The representative definitions are the definition of strong community and weak community [29]. Strong community requires that each node in the community have more links connecting to the nodes within the community than to the nodes outside the community. This definition is also called $LS$-set [30]. Accordingly, weak community requires that the sum of degrees for nodes within the community is larger than the number of links point to the rest of network. Note that a strong community is also a weak community. Hu et al. further gave an alternative definition to weak community and strong community by considering the links among different communities instead of considering the links between the target community and the rest of network [31]. The latter definition of strong and weak community is consistent with the constructing rule of benchmark networks [14] proposed by Girvan and Newman.

#### 1.3.1.2 Global Definitions to Community

Global definitions to community focus on the properties of the whole network rather than the properties of the community itself. The representative global definition is in terms of the network partition. By giving a measurement to evaluate the quality of network partition, we can find the optimal partition of network. This network partition provides the results of community, i.e., each component of the partition corresponds to a community. The well-known global definition to community is the modularity proposed by Girvan and Newman [32]. They take the configuration model as the null model to generate reference networks and characterize the modular structure of network by comparing the partition of real network with its randomized part in reference networks. The reference networks generated by the configuration model possess the same degree sequence to the real network. With the modularity at hand, the community structure can be detected by optimizing the modularity to find the optimal partition. The proposal of modularity greatly propels the development of community detection. Many optimization methods are then proposed to detect community by optimizing modularity with different heuristic strategies. Reichardt and Bornholdt considered the Potts model and gave an extended modularity [33]. Moreover, Rosvall and Bergstrom proposed a new measurement for network partition by studying the expected description length of a random walk on networks [34]. In addition, several probabilistic methods are proposed to model the network data. The likelihood of generating the network according to probabilistic model can be taken as an objective function which implicitly define the community of network [35].

As a summary, global definitions of community are accepted more widely than local definitions of community. The main reason lies in that global definition study community structure from the perspective of the whole network rather than from the perspective of the community itself. The basic idea behind each global definition to community corresponds to an insight to the community as a salient structural regularity of network.

## *1.3.2 Community Detection*

Community detection is the central research topic in network science. In the last decade, lots of literature devote to the detection of community structure in networks. Here, we will review the main development of community detection. For a thorough survey about community detection, readers can refer to Ref. [12].

### 1.3.2.1 Hierarchical Clustering

Community detection is closely related to the problem of graph clustering. Actually, most traditional methods for community detection are borrowed from graph clustering or graph cut [36]. Typical examples include the RatioCut [37] and NCut [38]. Among these traditional methods for community detection, the most successful methods are hierarchical clustering methods [39].

Hierarchical clustering methods can be classified into two classes: agglomerative methods and divisive methods. Agglomerative methods work in a bottom-up manner. At the beginning, each node is viewed as a community. Then communities are iteratively merged according to certain given measurement which quantifies the similarity between communities. The merging process terminates until all the nodes belong to the same community. On the contrary, divisive methods work in a top-down manner. At the beginning, all the nodes belong to the same community. Then we divide the community iteratively using certain given strategy until each node belongs to its own community. For both the agglomerative methods and the divisive methods, the final result of merging produces a dendrogram which depicts the merging sequence of communities. Cutting the dendrogram at any level, we obtain a partition of network. All the components of the network partition are viewed as the final communities.

Hierarchical clustering methods face two big challenges. The first one is the choice of measurement to determine the pair of communities to be merged in agglomerative methods or the community to be divided in divisive methods. Different choices of measurement lead to different hierarchical methods for community detection. In the seminal paper on community structure [14], edge betweenness is used as the measurement in a divisive method. Then, Newman and Girvan further propose three kinds of edge betweenness, which are calculated according to shortest path, random walk and current flow [32]. Furthermore, to avoid the high computational cost of edge betweenness, Radicchi et al. proposed edge-clustering coefficient as the measurement in divisive methods [29]. For agglomerative methods, Fortunato et al. proposed to use the information centrality as the measurement [40]. Moreover, the increase of modularity is widely used as the measurement in agglomerative methods [41, 42].

The second challenge is the choice of appropriate place to cut the dendrogram produced by hierarchical clustering methods. To combat this problem, we need a measurement to quantify the goodness of a network partition. The well-known measure is the modularity [32], which is proposed by Newman and Girvan when they study their divisive methods based on edge betweenness. The appearance of modularity greatly propels the development of community detection and motivates the prosperity of modularity optimization method.

#### 1.3.2.2 Modularity Optimization

Modularity is measurement to characterize modular property of network. Modularity is defined with respect to the partition of network. A partition with high modularity is viewed as a good partition in the sense that there are more edges within communities than expected.

Let $e_{st}$ be the fraction of edges in the network that connect nodes in group $s$ to those in group $t$ and $a_s$ is the fraction of all the *ends of edges* that are attached to nodes in group $s$. Then the modularity [32] is defined as

$$Q = \sum_s e_{ss} - a_s^2, \tag{1.1}$$

where $a_s = \sum_t e_{st}$ and $e_{ss}$ is the fraction of edges which connect nodes in community $s$. Note that $a_s^2$ is the fraction of edges that connect nodes within group $s$ when the ends of edges are connected at random. Then the physical meaning of the modularity $Q$ is clear: give a partition, the modularity is the difference between the real fraction of edges within communities and the expected fraction of edges within communities when edges are placed at random.

The modularity can also be defined on edges rather than being defined on communities (Eq. 1.1). This form of definition [42] can be written as

$$Q = \frac{1}{2m} \sum_{vw} \left[ A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w), \tag{1.2}$$

where $A_{vw}$ denotes whether there exists edge between nodes $v$ and $w$, $k_v = \sum_w A_{vw}$ is the degree of node $v$, $2m = \sum_v k_v$ is the total degree in the network, $c_v$ denotes the community node $v$ belongs to, and $\delta_{c_v, c_w}$ is 1 if $c_v = c_w$ and 0 otherwise. According this form of modularity, we can clearly see that the referenced null model for modularity is the configuration model, which is widely-used to generate random networks with the same degree sequence to the original network. Moreover, the modularity has been extended to weighted networks [41], directed networks [43, 44], bipartite networks [44, 45], and multiplex networks [46] by revising the referenced null model with different constraints.

With the modularity, we can find the optimal partition having the maximum modularity by searching the space composed of all the possible partitions of the given network. This optimal partition reveals the community structure of networks, i.e., each component of this partition is taken as one community. Therefore, the community structure of network can be detected by the optimization of modularity. Unfortunately, the optimization of modularity over all the possible partitions was proved to be a NP-hard problem [47]. Thus, scientists begin to find different heuristic optimization method to find the suboptimal partition with respect to the modularity. Examples include the greedy methods [41, 42, 48], the annealing method [20], the extremal optimization method [49], the spectral methods [50, 51], the genetic method [52], the mathematical programming method [53], the tabu-search method [54], and the multi-step optimization methods [55].

Modularity optimization method is the most widely used method for community detection. This kind of methods has gained great success on many specific real world networks. However, the optimization method suffers several severe problems. These problems pose big concerns on the reliability of the communities obtained by the modularity optimization. The first problem is about the referenced random network. Guimera et al. pointed out that the random networks generated by the configuration model can also exhibit high modularity due to the fluctuation of randomness [56]. To combat this problem, Sales-Pardo et al. proposed to use the statistical significance of modularity with respect to the networks generated by the referenced null model [57]. The second problem is the resolution limit problem [58] pointed out by Fortunato and Barthelemy. This problem tells us that the optimization of modularity depends on the intrinsic scale of network. For communities which are beyond this intrinsic scale cannot be detected by optimizing the modularity although these communities have clear identities. To overcome this problem, several multi-resolution methods are proposed using extended modularity [33, 54, 59]. The third problem for modularity optimization is that the results of modularity optimization are heavily affected by the degree distribution of network. Shen et al. proposed a rescaling remedy for this problem and can well deal with the networks with highly heterogeneous degree distribution [60]. In addition, as pointed out by Good et al., the modularity has a ragged landscape and thus the optimization of modularity exhibits extreme degeneracies: the optimization of modularity typically admits an exponential number of distinct high-scoring solutions and typically lacks a clear global maximum [61]. This problem further affects the practical performance of modularity.

#### 1.3.2.3 Network Dynamics

We have known that the modularity optimization method suffers several problems and these problems limit the applicability of this method. Meanwhile, researchers tried to understand the relationship between modularity and dynamic process on networks. Thus, dynamics on network provide another possible way to investigate the community structure [8, 9].

Arenas et al. studied the synchronization process on network and pointed out the synchronization process reveals the topological scales of network [62]. Multiple stable transients during the synchronization process correspond to multiscale community structure. They further pointed out the relationship between multiscale community structure and the spectrum of Laplacian matrix of network. By noticing that the synchronization process is affected by the heterogeneous degree distribution, Shen et al. studied the diffusion process on networks and reveals the relationship between the stable transients during diffusion process and the multiscale community structure [63]. Then, network conductance is used to detect the community structure associated with the diffusion dynamics on networks.

By investigating the random walk on networks, Rosvall et al. proposed the map equation to characterize the expected length of random walk path on networks [34]. With the map equation, people can find the optimal partition which gives the shortest description of random walk. The map equation method is well known as Infomap and is one of the most accurate partition-based community detection methods. Before the Infomap, Rosvall et al. investigates the relationship between network compression and community detection under the framework of information theory [64].

Raghavan et al. studied the label propagation process on networks and proposed the label propagation algorithm to detect community structure [65]. At the beginning of label propagation process, each node has a distinct label. Then each node replaces its label with the label which is most widely hosted label among its neighbors. Random selection is introduced to break a draw. When all the labels keep unchanged, the label propagation process terminates. Then, the nodes with the same label form one community. The label propagation process converges very fast and thus the label propagation algorithm is a near linear algorithm for community detection.

In summary, network dynamics is closely related to community structure of networks. It is critical to understand the structure and function of networks by revealing community structure by investigating network dynamics. We look forward to seeing more effective methods along this direction.

#### 1.3.2.4 Overlapping Community Detection

For all the above community detection methods, one node belongs to one and only one community. Because of such a constraint, the above methods cannot uncover the overlapping community structure. Actually, for real word networks, communities are highly overlapped. For example, in social networks, one person can simultaneously belong to multiple social circles, depending on his/her family, friends, professions, hobbies, and so on. In the network of words, several words have multiple meanings and thus belong to more than one communities of word. In scientific collaboration network, one researcher can collaborate with researchers from different research groups or even different countries. The nodes which participate in more than one community play crucial role in the function of networks. They bridge different communities and are the gatekeepers of their own communities. They facilitate the flow of information among different communities and are the potential factors for community evolution.

The detection of overlapping communities was first studied in Ref. [23]. In this seminal paper, a clique percolation method is described to uncover overlapping community structure. For clique percolation, a $k$-clique is rolled over the network to other cliques with $k-1$ common nodes. In this way, a community is composed of all the $k$-cliques which can reach each other by rolling on network. Because that one node can participate more than one $k$-cliques, it is likely that one node simultaneously belongs to multiple communities. The clique percolation method require us to find all the $k$-cliques in the network. This can be done in polynomial time complexity. However, in general, people are inclined to find all the maximal cliques, which is exponential in time complexity. In fact, finding all maximal cliques is more efficient than finding all the $k$-cliques since most real world networks are sparse. Kumpula et al. proposed a sequential algorithm for clique percolation [66]. This sequential algorithm greatly improves the computational efficiency of clique percolation. Furthermore, the clique percolation method is extended to weighted network [67], directed network [68], and bipartite networks [69]. The clique percolation methods have been successfully applied to biological networks, social networks and information networks. However, for networks with very few cliques, the percolation method is not applicable.

Many other methods are proposed to detect overlapping community structure in networks. Lancichinetti et al. detected overlapping communities by community expansion starting from different node seeds [70]. Baumes et al. gave a similar community expansion method with the difference at the choice of seed nodes [71]. Lee et al. developed an overlapping community detection method by expanding from maximal cliques instead of individual nodes [72]. By extending the label propagation algorithm designed for non-overlapping community detection, Gregory proposed the Community Overlap PRopagation Algorithm (COPRA) for overlapping community detection [73]. Gregory et al. also designed method to detect overlapping communities using extended edge betweenness [74]. Evans et al. proposed the line graph and detect communities which overlap at the level of nodes [75]. Furthermore, Ahn et al. investigated the link-communities and they clustered links to detect multiscale overlapping communities [76].

In recent years, probabilistic models attracted much research attention on overlapping community structure. This kind of methods focuses on the likelihood of observing the network according to certain probabilistic model. Then, by maximizing the likelihood, the parameters of model are determined and these parameters provide us information for overlapping community structure. Representative methods include the mixture model [35], the block models [77, 78] and the models based on latent Dirichlet allocation [79].

#### 1.3.2.5 Dynamic Community Detection

The above efforts for community detection are mainly devoted to the community structure of a static network. Actually, the structure of network evolves all the time. Accordingly, communities of network are also highly dynamic. However, the analysis of dynamic communities attracts only little attention. The main reasons are two-fold. Firstly, it is still controversial how to detect the intrinsic communities in static networks. The second reason lies in the difficulty of obtaining network data with time stamp. Recently, with the increasing availability of timestamped data, it is possible to study the evolution of communities, including the birth, growth, contraction, merge, split and death of communities.

The pioneering work on dynamic community detection is conducted by Hopcroft et al., who studied snapshots of the citation network obtained from the NEC CiteSeer Database [80]. To distinguish the real evolution of communities from the fluctuation on communities caused by the community detection methods, they only considered the *natural communities* defined as the communities only slightly affected by perturbations of network structure. Then they match the natural communities across different snapshots. In this way, they can track the evolution of communities in network. In 2007, Palla et al. systematically analyzed the dynamic communities in the network of mobile phone call and the scientific collaboration network [81]. They used the clique percolation method to detect the overlapping communities in different snapshot and compare the communities across different snapshots of network. They found that large communities persist longer if their memberships dynamically change while small communities keep stable if their members remain unchanged. Lin et al. performed the analysis of dynamic communities by considering both the quality of obtained communities in network snapshots and the consistency of community structure between successive network snapshots [82].

For dynamic community detection, most current work focuses on the snapshots of networks. This methodology may fail to uncover the mechanism of community evolution and predict the evolution of community. The evolution of explicit communities in cyberspace provide us important data for analyzing the mechanism of community evolution and designing algorithms to predict the evolution of these virtual communities.

## *1.3.3 Community Validity*

For the research on community structure, one critical problem is how to evaluate and compare the effectiveness of different methods for community detection. This problem includes two important aspects: how to construct the benchmark networks with known community structure and how to design the measurements to compare the known communities with the communities obtained by community detection methods.

### 1.3.3.1 Benchmark Networks

Benchmark networks are used to compare the performance of different methods. In the benchmark networks, the communities are known a priori. Thus, we need to plant the communities into a network without communities. Furthermore, we need parameters to control the ambiguity of community structure in benchmark networks. Here, we describe two kinds of well-known benchmark networks: the GN benchmark networks [32] proposed by Girvan and Newman and the LFR benchmark [83] networks proposed by Lancichinetti, Fortunato, and Radicchi.

The GN benchmark is called the standard benchmark since it is widely-used as the standard test for community detection methods. This benchmark is a realization of the so-called planted $l$-partition model [84]. For a network generated by the planted $l$-partition model, all the nodes are partitioned into $l$ groups and each group has $g$ nodes. Nodes of the same group are linked with a probability $p_{in}$ while nodes of different groups are linked with probability $p_{out}$. Note that each node group spans a network generated according to Erdős-Rényi random graph model [6] with the link probability $p_{in}$. The average degree of each node is $\langle k \rangle = p_{in}(g-1) + p_{out}g(l-1)$. If $p_{in} \lg p_{out}$, the intra-group link density is larger than the inter-group link density and we say that the generated benchmark networks have a community structure, i.e., each group is a community. The GN benchmark is a special case the planted $l$-partition model. In GN benchmark, $l = 4$, $g = 32$ and the average node degree $\langle k \rangle = 16$. According to these three constraints, the probability $p_{in}$ and the probability $p_{out}$ are not independent. In fact, we have $p_{in} + p_{out} \approx 1/2$. For convenience, it is common to use parameters $z_{in} = p_{in}(g-1) = 31p_{in}$ and $z_{out} = p_{out}g(l-1) = 96p_{out}$. The parameter $z_{in}$ is the expected internal degree of a node and $z_{out}$ is the expected external degree of a node. Note that $z_{in} + z_{out} = 16$. Generally, the communities in the benchmark network are well-defined when $z_{out} < 8$. In this situation, the communities satisfy the definition of strong community [29]. When $z_{out} > 8$, the definition of strong community is violated and we say that the benchmark network has no intrinsic community structure. Fan et al. gave a weighted version of the GN benchmark [85]. Furthermore, Arenas et al. proposed the hierarchical version of the GN benchmark [62] and Sawardecker et al. extend this benchmark to allow the communities overlap with each other [86].

The GN benchmark plays crucial role in the development of community detection. However, for GN benchmark, the node degree and the community size are both homogeneous. This is not consistent with the real world networks which exhibit heterogeneous distributions of node degree and community size. To combat the shortcoming of the GN benchmark, Lancichinetti et al. proposed the LFR benchmark [83]. In networks generated by the LFR benchmark, the node degree follows a power law distribution with exponent $\gamma$ and the community size follows another power law distribution with exponent $\beta$. In addition, a mixing ratio parameter $\mu$ is used to control the ambiguity of community structure in networks. Intuitively, for each node, $\mu$ denotes the average fraction of links pointing to the nodes of the other communities. The larger the parameter $\mu$ is, the more ambiguity the community structure is. In sum, the LFR benchmark poses more severe test to community detection methods than the GN benchmark because of the heterogeneous distributions of node degree and community size. Moreover, Lancichinetti et al. extended the LRF benchmark to directed, weighted networks and communities are allowed to overlap with each other [87].

Besides the synthetic benchmark networks, several real world networks are often used to test the effectiveness of community detection methods. These networks include the karate club network first studied by Zachary [88], the network of bottlenose dolphins living in New Zealand [89], the network of 115 football teams of

American Universities [14], and the citation networks of papers published on Physical Review Letters [64].

#### 1.3.3.2 Measurements

Measurements for community detection quantify the results of community detection methods. Here, we introduce three widely-used measurements for community detection.

The first measurement is the fraction of nodes identified accurately by community detection methods [14]. This simple measure is biased if the two communities are merged by the community detection methods. To combat this problem, normalized mutual information is proposed [90]. Both the obtained communities and the known communities are taken as partitions of network. Normalized mutual information compares the similarity between these two partitions. We denote the two partitions as $\mathscr{X} = (X_1, X_2, \ldots, X_{n_X})$ and $\mathscr{Y} = (Y_1, Y_2, \ldots, Y_{n_Y})$. Here, $n_X$ and $n_Y$ respectively denote the number of communities associated with the two partitions. In addition, we denote by $n$ the number of nodes in network, by $n_i^X$ the number of nodes in the community $X_i$, by $n_j^Y$ the number of nodes in the community $Y_j$, and by $n_{ij}$ the number of nodes shared by communities $X_i$ and $Y_j$. We use a variable $X$ to denote the community label of a randomly selected node according to the partition $\mathscr{X}$ and use a variable $Y$ to denote the community label of a randomly selected node according to the partition $\mathscr{Y}$. Then, the joint distribution $P(X_i, Y_j) = P(X = X_i, Y = Y_j) = n_{ij}/n$. Thus, $P(X_i) = P(X = X_i) = n_i^X/n$ and $P(Y_i) = P(Y = Y_i) = n_i^Y/n$. Then, the normalized mutual information is defined as

$$NMI(\mathscr{X}, \mathscr{Y}) = \frac{2I(X,Y)}{H(X) + H(Y)}, \tag{1.3}$$

where the mutual information $I(X,Y) = H(X) - H(X|Y)$, then Shannon entropy of $X$ is $H(X) = -\sum_i P(X_i) \log P(X_i)$ and the conditional entropy of $X$ given $Y$ is $H(X|Y) = -\sum_{ij} P(X_i, Y_j) \log P(X_i|Y_j)$. The normalized mutual information is 1 is the two partitions $\mathscr{X}$ and $\mathscr{Y}$ are identical and 0 if the two partitions are independent. The larger the normalized mutual information is, the more similar the two partitions are.

As an alternative to the normalized mutual information, Meilă introduced the variation of information [91], which is defined as

$$Var(\mathscr{X}, \mathscr{Y}) = H(X|Y) + H(Y|X). \tag{1.4}$$

Compared to the normalized mutual information, the variation of information defines a distance in the space of partitions. However, since the maximum of the variation of information is $\log n$, the value of variation of information for networks with different sizes cannot be compared with each other. For comparison, one could use the normalized version, i.e., $Var(\mathscr{X}, \mathscr{Y})$ divided by $\log n$ as suggested by Karrer et al. [92].

## 1.4 Concluding Remarks

In this chapter, we have introduced the background of network science and the development in community detection. We know that community structure is a salient structural characteristic of most real world networks. The community detection has attracted much research attention from various fields. However, several open problems still exists and these problems provide the motivations of this book. These problems include: 1) How to quantify the overlapping community structure in networks and how to simultaneously uncover the overlapping and hierarchical community structure of networks? 2) How to detect the multiscale community structure in networks with heterogeneous degree distribution? 3) How to uncover the community structure associated with network dynamics? 4) How to infer the latent community structure according to observed links of networks, with only positive links or with both positive and negative links? In the remaining parts, we will introduce our research works on these problems.

## References

1. Strogatz, S.H.: Exploring complex networks. Nature **410**, 268–276 (2001)
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
3. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
4. Euler, L.: Solutio problematis ad geometrian situs pertinentis. Comment. Acad. Sci. Petropolitanae **8**, 128–140 (1736)
5. Bollobas, B.: Modern Graph Theory. Springer, New York (1998)
6. Erdős, P., Rényi, A.: On random graphs. Publ. Math. Debrecen **6**, 290–297 (1959)
7. Mendes, J.F.F., Dorogovtsev, S.N.: Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford (2003)
8. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Phys. Rep. **424**, 175–308 (2006)
9. Barrat, A., Barthélemy, M., Vespignani, A.: Dynamical Processes on Complex Networks. Cambridge University Press, Cambridge (2008)
10. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**, 509–512 (1999)
11. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**, 440–442 (1998)
12. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
13. Leskovec, J., Lang, K.J., Mahoney, M.W.: Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International Conference on World Wide Web, pp. 631–640 (2010)
14. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**, 7821–7826 (2002)
15. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA **98**, 404–409 (2001)
16. Freeman, L.C.: The Development of Social Network Analysis: A Study in the Sociology of Science. BookSurge Publishing, North Charleston (2004)
17. Flake, G.W., Lawrence, S.R., Giles, C.L., Coetzee, F.M.: Self-organization and identification of Web communities. IEEE Comput. **35**, 66–71 (2002)

18. Rives, A.W., Galitski, T.: Modular organization of cellular networks. Proc. Natl. Acad. Sci. USA **100**, 1128–1133 (2003)
19. Chen, J., Yuan, B.: Detecting functional modules in the yeast protein interaction network. Bioinformatics **22**, 2283–2290 (2006)
20. Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Nature **433**, 895–900 (2005)
21. Williams, R.J., Martinez, N.D.: Simple rules yield complex food webs. Nature **404**, 180–183 (2000)
22. Krawczyk, M.J.: Differential equations as a tool for community identification. Phys. Rev. E **77**, 065701 (2008)
23. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)
24. Alba, R.D.: A graph-theoretic definition of a sociometric clique. J. Math. Sociol. **3**, 113–126 (1973)
25. Mokken, R.J.: Cliques, clubs and clans. Qual. Quant. **13**, 161–173 (1979)
26. Seidman, S.B., Foster, B.L.: A graph theoretic generalization of the clique concept. J. Math. Sociol. **6**, 139–154 (1978)
27. Seidman, S.B.: Network structure and minimum degree. Soc. Netw. **5**, 269–287 (1983)
28. Matsuda, H., Ishihara, T., Hashimoto, A.: Classifying molecular sequences using a linkage graph with their pairwise similarities. Theoret. Comput. Sci. **210**, 305–325 (1999)
29. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA **101**, 2658–2663 (2004)
30. Borgatti, S.P., Everett, M.G., Shirey, P.: LS sets, lambda sets and other cohesive subsets. Soc. Netw. **12**, 337–357 (1990)
31. Hu, Y., Chen, H., Zhang, P., Li, M., Di, Z., Fan, Y.: Comparative definition of community and corresponding identifying algorithm. Phys. Rev. E **78**, 026121 (2008)
32. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
33. Reichardt, J., Bornholdt, S.: Detecting fuzzy community structures in complex networks with a Potts model. Phys. Rev. Lett. **93**, 218701 (2004)
34. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA **105**, 1118–1123 (2008)
35. Newman, M.E.J., Leicht, E.A.: Mixture models and exploratory analysis in networks. Proc. Natl. Acad. Sci. USA **104**, 9564–9569 (2007)
36. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell Syst. Tech. J. **49**, 291–307 (1970)
37. Wei, Y.C., Cheng, C.K.: Toward efficient hierarchical designs by ratio cut partitioning. In: Proceedings of IEEE International Conference on Computer Aided Design, pp. 298–301 (1989)
38. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 888–905 (2000)
39. Hastie, T., Tibshirani, R., Friedman, J.H.: The elements of statistical learning. Springer, Berlin (2001)
40. Fortunato, S., Latora, V., Marchiori, M.: Method to find community structures based on information centrality. Phys. Rev. E **70**, 056104 (2004)
41. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E **69**, 066133 (2004)
42. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**, 066111 (2004)
43. Leicht, E.A., Newman, M.E.J.: Community structure in directed networks. Phys. Rev. Lett. **100**, 118703 (2008)
44. Guimerà, R., Sales-Pardo, M., Amaral, L.A.N.: Module identification in bipartite and directed networks. Phys. Rev. E **76**, 036102 (2007)
45. Barber, M.J.: Modularity and community detection in bipartite networks. Phys. Rev. E **76**, 066102 (2007)

46. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community structure in time-dependent, multiscale, and multiplex networks. Science **328**, 876–878 (2010)
47. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Trans. Knowl. Data Eng. **20**, 172–188 (2008)
48. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. P10008 (2008)
49. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Phys. Rev. E **72**, 027104 (2005)
50. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA **103**, 8577–8582 (2006)
51. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**, 036104 (2006)
52. Tasgin, M., Herdagdelen, A., Bingol, H.: Community detection in complex networks using genetic algorithms (2007). arXiv:0711.0491
53. Agarwal, G., Kempe, D.: Modularity-maximizing graph communities via mathematical programming. Eur. Phys. J. B **66**, 409–418 (2008)
54. Arenas, A., Fernández, A., Gómez, S.: Analysis of the structure of complex networks at different resolution levels. New J. Phys. **10**, 053039 (2008)
55. Schuetz, P., Caflisch, A.: Multistep greedy algorithm identifies community structure in real-world and computer-generated networks. Phys. Rev. E **78**, 026112 (2008)
56. Guimerà, R., Sales-Pardo, M., Amaral, L.A.N.: Modularity from fluctuations in random graphs and complex networks. Phys. Rev. E **70**, 025101 (2004)
57. Sales-Pardo, M., Guimerà, R., Moreira, A.A., Amaral, L.A.N.: Extracting the hierarchical organization of complex systems. Proc. Natl. Acad. Sci. USA **104**, 15224–15229 (2007)
58. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proc. Natl. Acad. Sci. USA **104**, 36–41 (2007)
59. Ronhovde, P., Nussinov, Z.: Multiresolution community detection for megascale networks by information-based replica correlations. Phys. Rev. E **80**, 016109 (2009)
60. Shen, H.W., Cheng, X.Q., Fang, B.X.: Covariance, correlation matrix, and the multiscale community structure of networks. Phys. Rev. E **82**, 016114 (2010)
61. Good, B.H., Montjoye, Y., Clauset, A.: Performance of modularity maximization in practical contexts. Phys. Rev. E **81**, 046106 (2010)
62. Arenas, A., Díaz-Guilera, A., Pérez-Vicente, C.J.: Synchronization reveals topological scales in complex networks. Phys. Rev. Lett. **96**, 114102 (2006)
63. Cheng, X.Q., Shen, H.W.: Uncovering the community structure associated with the diffusion dynamics on networks. J. Stat. Mech. P04024 (2010)
64. Rosvall, M., Bergstrom, C.T.: An information-theoretic framework for resolving community structure in complex networks. Proc. Natl. Acad. Sci. USA **104**, 7327–7331 (2007)
65. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**, 036106 (2007)
66. Kumpula, J.M., Kivelä, M., Kaski, K., Saramäki, J.: Sequential algorithm for fast clique percolation. Phys. Rev. E **78**, 026109 (2008)
67. Farkas, I.J., Ábel, D., Palla, G., Vicsek, T.: Weighted network modules. New J. Phys. **9**, 180 (2007)
68. Palla, G., Farkas, I.J., Pollner, P., Derényi, I., Vicsek, T.: Directed network modules. New J. Phys. **9**, 186 (2007)
69. Lehmann, S., Schwartz, M., Hansen, L.K.: Biclique communities. Phys. Rev. E **78**, 016108 (2008)
70. Lancichinetti, A., Fortunato, S., Kertesz, J.: Detecting the overlapping and hierarchical community structure of complex networks. New J. Phys. **11**, 033015 (2009)
71. Baumes, J., Goldberg, M., Magdon-Ismail, M.: Efficient identification of overlapping communities. Lect. Notes Comput. Sci. **3495**, 27–36 (2005)
72. Lee, C., Reid, F., McDaid, A., Hurley, N.: Detecting highly overlapping community structure by greedy clique expansion. In: Proceedings of the 4th SNA-KDD Workshop (2010)

73. Gregory, S.: Finding overlapping communities in networks by label propagation. New J. Phys. **12**, 103018 (2010)
74. Gregory, S.: An algorithm to find overlapping community structure in networks. In: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 91–102 (2007)
75. Evans, T.S., Lambiotte, R.: Line graphs, link partitions, and overlapping communities. Phys. Rev. E **80**, 016105 (2009)
76. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**, 761–764 (2010)
77. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. J. Mach. Learn. Res. **9**, 1981–2014 (2008)
78. Karrer, B., Newman, M.E.J.: Stochastic blockmodels and community structure in networks. Phys. Rev. E **83**, 016107 (2011)
79. Eorsheva, E., Fienberg, S., Lafferty, J.: Mixed-membership models of scientific publications. Proc. Natl. Acad. Sci. USA **101**, 5220–5227 (2004)
80. Hopcroft, J., Khan, O., Kulis, B., Selman, B.: Tracking evolving communities in large linked networks. Proc. Natl. Acad. Sci. USA **101**, 5249–5253 (2004)
81. Palla, G., Barabási, A.L., Vicsek, T.: Quantifying social group evolution. Nature **446**, 664–667 (2007)
82. Lin, Y.R., Chi, Y., Zhu, S.H., Sundaram, H., Tseng, B.L.: FacetNet: A framework for analyzing communities and their evolutions in dynamic networks. In: Proceedings of the 17th International Conference on World Wide Web, pp. 685–694 (2008)
83. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**, 046110 (2008)
84. Condon, A., Karp, R.M.: Algorithms for graph partitioning on the planted partition model. Random Struct. Algorithms **18**, 116–140 (2001)
85. Fan, Y., Li, M., Zhang, P., Wu, J., Di, Z.: Accuracy and precision of methods for community identification in weighted networks. Physica A **377**, 363–372 (2007)
86. Sawardecker, E.N., Sales-Pardo, M., Amaral, L.A.N.: Detection of node group membership in networks with group overlap. Eur. Phys. J. B **67**, 277–284 (2009)
87. Lancichinetti, A., Fortunato, S.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E **80**, 016118 (2009)
88. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**, 452–473 (1977)
89. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait? Behav. Ecol. Sociobiol. **54**, 396–405 (2003)
90. Danon, L., Duch, J., Diaz-Guilera, A., Arenas, A.: Comparing community structure identification. J. Stat. Mech. P09008 (2005)
91. Meilă, M.: Comparing clusterings: An information based distance. J. Multivar. Anal. **98**, 873–895 (2007)
92. Karrer, B., Levina, E., Newman, M.E.J.: Robustness of community structure in networks. Phys. Rev. E **77**, 046119 (2008)

# Chapter 2
# Detecting the Overlapping and Hierarchical Community Structure in Networks

## 2.1 Introduction

As described in the previous chapter, community structure is a common and important topological characteristic of many real world complex networks. Examples include the World Wide Web, citations networks, various kinds of social and biological networks, and many others [1–3]. In the past decade, community structure has attracted much research attention from various scientific fields since it is crucial to understand the structural and functional properties of networks [4–6]. Many methods have been proposed to identify the community structure of complex networks [7–13]. The reader can refer to Ref. [14] for reviews.

These existing methods can be roughly classified into two categories in terms of the form of their results, i.e., to form a partition or a cover of the network. The first kind of methods produce a partition, i.e., each node belongs to one and only one community and is regarded as equally important. Different from classical graph partition problem, the number of communities and the size of each community are prior unknown. Among this kind of methods, the most successful ones are the methods based on the optimization of modularity [11, 15, 16], which is proposed by Newman et al. as a quality function to measure the goodness of a network partition [9]. A high value of modularity indicates a significant community structure. Generally, this kind of methods is suitable to understand the community structure of the whole networks, especially for the networks with small sizes. However, the modularity optimization methods also suffer several problems, e.g., the resolution limit problem [17, 18]. These problems pose concerns about the reliability of the community structure detected by directly optimizing the modularity.

The second kind of methods aim to discover the node sets i.e., *communities* with a high density of edges. In this case, overlapping is allowed, that is, some nodes may belong to more than one community. Meanwhile, some nodes may be neglected as subordinate nodes. Therefore, these methods result in an incomplete cover of the network. Compared to the partition methods, this kind of methods are appropriate to find the cohesive regions in the large scale networks. Ever since the problem of detecting overlapping community structure is proposed by Palla et al., many meth-

H.-W. Shen, *Community Structure of Complex Networks*, Springer Theses,
DOI 10.1007/978-3-642-31821-4_2, © Springer-Verlag Berlin Heidelberg 2013

ods have been proposed [8, 19–26]. In [8], the community structure is uncovered by $k$-clique percolation and the overlaps between communities are guaranteed by the fact that one node can participate in more than one clique. However, the $k$-clique method gives rise to an incomplete cover of network, i.e., some nodes may not belong to any community. In addition, the hierarchical structure cannot be revealed for a given $k$. In [24], by introducing the concept of the belonging coefficients of each node to its communities, the authors proposed a general framework for extending the traditional modularity to quantify overlapping community structure. The method provides a new idea to find overlapping community structure. However, the physical meaning of the belonging coefficient lacks a clear explanation. Furthermore, the framework is hard to extend to large scale networks since it is difficult to find an efficient algorithm to search the huge solution space. Recently, Evans et al. [25] proposed a method to identify the overlapping community structure by partitioning a line graph constructed from the original network. This method only allows the communities to overlap at nodes. More important, there is no commonly accepted standard to evaluate the goodness of a cover up to now.

In real networks, communities are usually overlapping and hierarchical [8, 26–29]. Overlapping means that some nodes may belong to more than one community. Hierarchical means that communities may be further divided into sub-communities. The two kinds of existing methods, as mentioned above, investigate these two phenomena separately. The first kind of methods can be used to explore the hierarchical community structure. However, they are unable to deal with overlaps between communities. The second kind of methods can uncover overlapping community structure of networks, but they are incapable of finding the hierarchy of communities. Recently, Lancichinetti et al. make a pioneering attempt on the detection of both overlapping and hierarchical community structure in complex networks [26]. They try to detect the overlapped communities in the network based on the local optimization of a fitness function. Their method can uncover the hierarchical relation between these overlapped communities around a particular node. The remained problem lies in that the detection of the hierarchy of all overlapped communities in the network is not guaranteed due to the random choice of seed nodes.

In this chapter, we focus on the problem of detecting the overlapping and hierarchical community structure simultaneously. By taking maximal cliques as basic building blocks of communities, we propose an algorithm EAGLE (agglomerativE hierarchicAl clusterinG based on maximaL cliquE) to detect community structure of networks. The overlaps among different maximal cliques guarantee the overlaps between communities and the hierarchy of these overlapped communities is uncovered by the process of agglomerative hierarchical clustering. Then, by extending the measurement for network partition, i.e., modularity, we propose a measurement $Q_c$ to evaluate the quality of a cover of network. Then, we can find the overlapping community structure by directly optimizing the proposed measurement. Furthermore, we propose a method to construct a maximal clique network for a given network. With the maximal clique network at hand, finding the overlapping community structure by optimizing the new measurement $Q_c$ on the original networks is equivalent to optimizing the standard modularity on the maximal clique network. In this way, any method based on modularity optimization can be directly used to uncover the overlapping community structure.

## 2.2 EAGLE: Detecting the Overlapping and Hierarchical Community Structure

In this chapter, the algorithm EAGLE is presented to uncover both the overlapping and hierarchical community structure of networks. This algorithm deals with the set of maximal cliques and adopts an agglomerative framework. The effectiveness is then demonstrated by applications to two real world networks, namely the word association network and the scientific collaboration network.

### *2.2.1 The Algorithm*

Before we introduce the details of the algorithm EAGLE, we use a schematic network to illustrate what EAGLE can do and compare it with the representative algorithms of the two kinds of existing methods introduced in the previous section. Figure 2.1(a) depicts the schematic network. This network is constructed according to the schematic network in [8], which has overlapping community structure. To construct the hierarchy of the overlapped communities, we remove the edge connecting nodes 9 and 13 and add two edges, one connecting 10 and 15 and the other one connecting 10 and 13. Figure 2.1(b) shows the community structure found by Newman's fast algorithm [11], which is the representative algorithm of partitioning network based on modularity optimization. Three communities are found when applying the algorithm to the schematic network. The hierarchy of communities can be revealed by applying the algorithm to each community further. For example, one of the three communities is divided into two sub-communities. However, overlaps between communities are not allowed. Figure 2.1(c) demonstrates the overlapping community structure found by the $k$-clique algorithm [8], which is the representative algorithm producing a cover of network. Unfortunately, this algorithm cannot reveal the hierarchy of community. Figure 2.1(d) shows the hierarchical and overlapping community structure found by the algorithm EAGLE. We can see that the algorithm EAGLE provides a possible way to investigate a more complete picture of the community structure.

Now we turn to the basic ideas behind the algorithm EAGLE. Generally speaking, a community can be regarded as a node set within which the nodes are more likely connected to each other than to the rest of the network. This indicates that a community usually has relatively high link-density. We know that the link-density of a clique is highest among all kinds of node subsets of a network. Furthermore, a densely-linked community usually contains a large clique, which could be regarded as the core of the community. Based on this observation, the algorithm EAGLE is proposed as an agglomerative hierarchical clustering algorithm to investigate

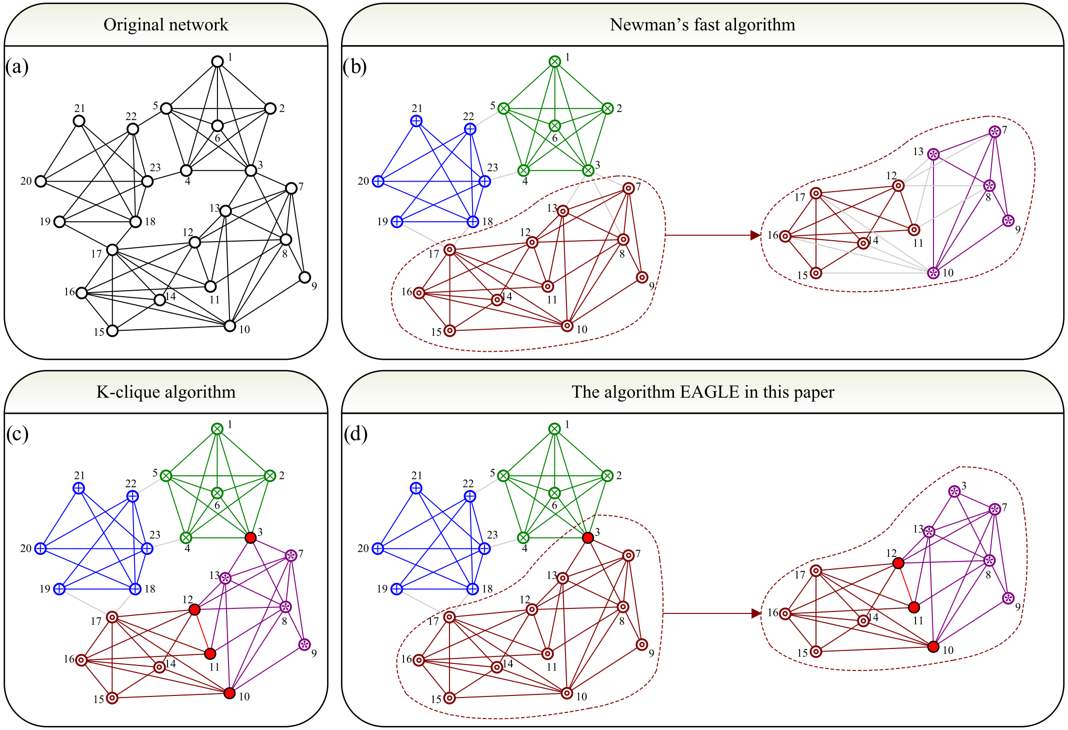


**Fig. 2.1** Comparison of community structure found by different algorithms. Different communities are rendered in different colors (or markers for print). Edges between communities are colored in *light gray*. Overlapping regions between communities are emphasized in *red*. (**a**) The schematic network. (**b**) The hierarchical community structure found by Newman's fast algorithm. This algorithm is chosen as a representative of the first kind of algorithms. (**c**) The overlapping community structure found by the *k*-clique algorithm as a representative of the second kind of algorithms. (**d**) The hierarchical and overlapping community structure found by the algorithm EAGLE. Reprinted from Ref. [30], Copyright 2009, with permission from Elsevier

the community structure. Different from traditional agglomerative algorithms [11], the algorithm deals with the set of maximal cliques rather than the set of nodes.

A *maximal clique* is a clique which is not a subset of any other cliques. In the algorithm EAGLE, we need to first find out all the maximal cliques in the network. This can be done by many efficient parallel algorithms. Here we choose the well-known *Bron-Kerbosch* algorithm [31] for its efficiency and its simplicity in implementation. Note that not all maximal cliques are taken into account. The maximal cliques, whose nodes are from some other larger maximal cliques, are called *subordinate maximal cliques*. For example, in Fig. 2.1, nodes 4 and 23 form a subordinate maximal clique. Because node 4 is from another larger maximal clique $\{1, 2, 3, 4, 5, 6\}$ and node 23 is also from other larger maximal cliques, including $\{18, 20, 21, 23\}$, $\{18, 20, 22, 23\}$ and $\{18, 19, 22, 23\}$. Subordinate maximal cliques may mislead our algorithm and thus are discarded. Most subordinate maximal cliques have small sizes. Thus, we can discard them by setting a threshold $k$ and neglecting all the maximal cliques with the size smaller than $k$. This simple tactic may also discard some non-subordinate maximal cliques. The higher the value of $k$ is, the more non-subordinate maximal cliques are discarded by mistake. On the other hand, the smaller the value of $k$ is, the more subordinate maximal cliques are remained. In real world networks, the threshold $k$ typically takes value between 3 and 6. As to the network in Fig. 2.1, both 3 and 4 are appropriate threshold values. As to the networks used in Sect. 2.2.2, 4 is demonstrated to be an appropriate threshold [8]. After neglecting the maximal clique with the size smaller than the threshold $k$, some nodes do not belong to any remaining maximal cliques. We call these nodes as *subordinate nodes*.

The algorithm EAGLE has two stages. In the first stage, a dendrogram is generated. In the second stage, we choose an appropriate cut which breaks the dendrogram into communities. The first stage of the algorithm EAGLE can be described as follows:

1. Find out all maximal cliques in the network. Neglect subordinate maximal cliques. The remainders are taken as the initial communities. Each subordinate node is also taken as an initial community comprising the sole node. Calculate the similarity between each pair of communities.
2. Select the pair of communities with the maximum similarity, incorporate them into a new one and calculate the similarity between the new community and other communities.
3. Repeat step 2 until only one community remains.

In the algorithm, the similarity $S$ between two communities $C_1$ and $C_2$ is defined as

$$S = \frac{1}{2m} \sum_{v \in C_1, w \in C_2, v \neq w} \left[ A_{vw} - \frac{k_v k_w}{2m} \right], \tag{2.1}$$

where $A_{vw}$ is the element of adjacency matrix of the network (here, we only consider undirected, unweighted networks). It takes value 1 if there is an edge between node $v$ and node $w$ and 0 otherwise. The quantity $m = \frac{1}{2} \sum_{vw} A_{vw}$ is the total number of edges in the network and $k_v$ is the degree of node $v$.

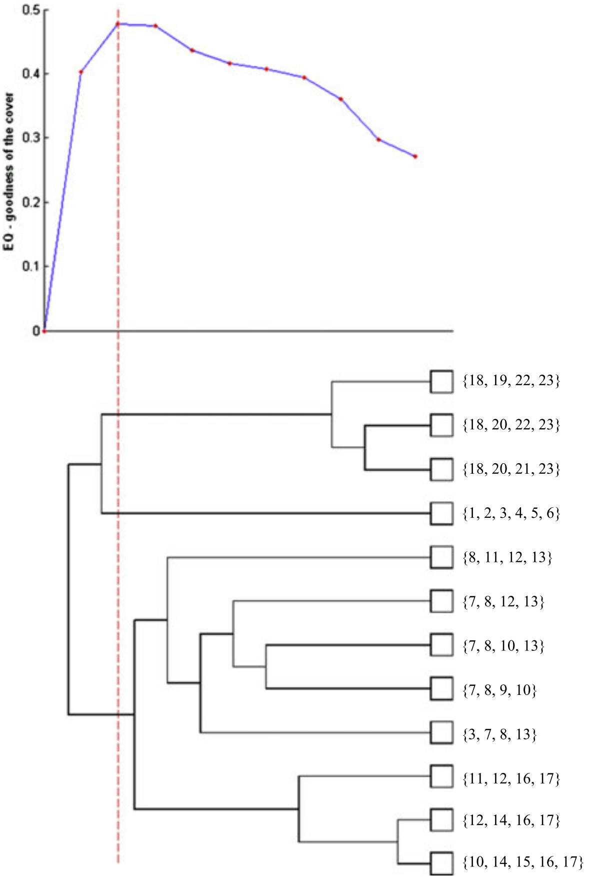


**Fig. 2.2** Illustration for the process of the algorithm EAGLE. This illustration is according to the schematic network in Fig. 2.1. The *bottom part* is a dendrogram. The leaf nodes correspond to the non-subordinate maximal cliques. The label of each leaf node shows the nodes belonging to it. The (*red*) *vertical dashed line* is a cut through the dendrogram and it gives the best cover of the network. The *top part* of the figure is a graph which illustrates the curve of *EQ* corresponding to each cover of the network. The threshold *k* is set to be 4. Reprinted from Ref. [30], Copyright 2009, with permission from Elsevier

Similar to the fast algorithm in [11], the process of our algorithm corresponds to a dendrogram, which shows the order of the amalgamations of communities. Any cut through the dendrogram produces a cover of the network. As an illustration, Fig. 2.2 shows the dendrogram generated by our algorithm when applied to the network in Fig. 2.1.

The task of the second stage of the algorithm EAGLE is to cut the dendrogram. To determine the place of the cut, a measurement is required to judge the quality of a cover. In [24], an extension of modularity is proposed to evaluate the goodness of overlapped community decomposition. Here, for simplicity, we propose another extension of modularity, namely *EQ*. In Fig. 2.2, the cut gives the best cover with the maximum value of *EQ*. The extended modularity is defined as

$$EQ = \frac{1}{2m}\sum_{i}\sum_{v\in C_i, w\in C_i}\frac{1}{O_v O_w}\left[A_{vw} - \frac{k_v k_w}{2m}\right], \tag{2.2}$$

where $O_v$ is the number of communities to which node $v$ belongs.

Note that *EQ* reduces to $Q$ in [9] when each node belongs to only one community (readers can refer to Ref. [15] for details), and *EQ* is equal to 0 when all nodes belong to the same community. In addition, it will be shown later in Sect. 2.2.2, a high value of *EQ* indicates a significant overlapping community structure.

Alike to modularity, the extended modularity suffers a resolution limit beyond which no modular structure can be detected even though these modules might have their own identities. For the algorithm EAGLE, however, these modules can be still detected by further applying the algorithm to each community found until none of them can be divided into smaller ones. Thus, we obtain a hierarchy of overlapping communities which reveals the community structure of network more completely.

Now we analyze the time complexity of the algorithm. Let $n$ be the number of nodes, $s$ be the number of maximal cliques in the initial state of the algorithm, and $h$ be the number of pair of maximal cliques which are neighbors (connected by edges or overlap with each other). We firstly consider the first stage of the algorithm. In step 1, $O(n^2)$ operations are needed to calculate the similarity between each pair of initial communities. In step 2, we only consider the pairs of communities which are neighbors. Each selection costs $h$ operations and each time of join costs $O(n)$ operations at most. Totally, we carry on a maximum of $s - 1$ join operations. Thus the first stage of the algorithm takes at most $O(n^2 + (h + n)s)$ operations. As to the second stage, we need to calculate the value of *EQ* corresponding to each cover. In our implementation, we calculate the value of *EQ* for the initial cover and update it after each join of two selected communities into a new one. Each time of update costs at most $n^2$ operations. Hence, the second state of the algorithm takes at most $O(n^2s)$ operations. In addition, we need to find out all the maximal cliques in the network. It is widely believed to be a non-polynomial problem. However, for real world networks, finding all the maximal cliques is easy due to the sparseness of these networks. Compared to the Newman's fast algorithm and the $k$-clique algorithm, the algorithm EAGLE is time-consuming. We leave it as a future work that how to improve the speed of EAGLE.

### 2.2.2 Applications

In this subsection, we apply the algorithm EAGLE to two real world complex networks, the word association network and the scientific collaboration network. The results show that EAGLE can discover new knowledge and insights underlying these networks.

The test data of the two networks are from the demo of the *CFinder* [32].[1] The two networks comprise $7207, 16662$ nodes and $31784, 22446$ edges, respectively. The average clustering coefficients [33] are approximately 0.15 and 0.19, which indicate that these networks have significant community structures in general.

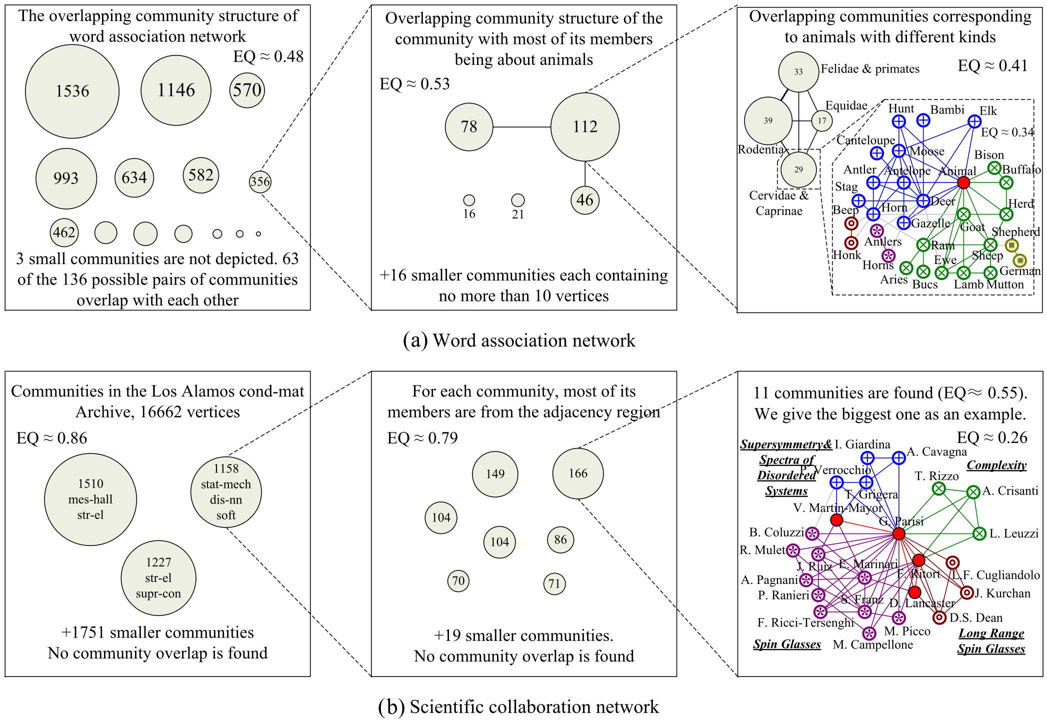
[1] CFinder is a free software for finding overlapping dense groups of nodes in networks, based on the clique percolation method. URL: www.cfinder.org.

The word association network is constructed from the South Florida Free Association norms list. The original network is directed and weighted. The weight of a directed link from one word to another indicates the frequency that the people in the survey associated the end point of the link with its start point. The directed links are replaced by undirected ones with a weight equal to the sum of the weights of the corresponding two oppositely directed links. Furthermore, the links with the weight less than 0.025 are deleted. The scientific collaboration network is from the co-authorship network of *Los Alamos e-print* archives. Each article in the archive between April 1998 and February 2004 contributes the value $1/(n-1)$ to the weight of the link between every pair of its $n$ authors. The link with the weight less than 1.0 is omitted.

In the word association network, totally 17 communities are found by our algorithm—see Fig. 2.3(a), left panel. Among these communities, 63 of 136 possible pairs of communities overlap with each other. To investigate what is correlated to the community structure, we apply our algorithm to each of these communities again. The sub-community structure of one community is given in Fig. 2.3(a), middle panel. Each of these sub-communities has certain correlation with the semantic meaning of words. For example, most of the words in the community with size 112 are related to the family of animals in Africa. This community is explored further and four communities are found, shown in Fig. 2.3(a), right panel. Each community is associated with animals from the same family, namely rodentia, felidae & primates, cervidae & caprinae, and equidae respectively. The details of one community are also illustrated in Fig. 2.3(a), right panel. Two large communities correspond to words associated with animals from cervidae and caprinae respectively. The overlapped word *Animal* acts as a bridge between the two communities. Three small communities comprise peripheral words.

Applying our algorithm to the scientific collaboration network, we obtain totally 1754 communities—see Fig. 2.3(b), left panel, with the corresponding high value of $EQ \approx 0.86$. Three large communities contain 23.4 % of all the nodes, while the others are relatively small. The three large communities correspond closely to subject subareas: the biggest one mainly to *mes-hall* and *str-el*, the second biggest one to *str-el* and *supr-con*, and the other to *stat-mech*, *dis-nn* and *soft*. We further apply the algorithm to one community and it is broken down into 26 sub-communities—depicted in Fig. 2.3(b), middle panel. There appears to be a correlation between the sub-community structure and the regional divisions of the scientific researchers. For example, most of the members of the community with size 166 work in Europe. More specific regional information can be obtained when applying the algorithm to this community. The biggest one and its sub-community structure are given in Fig. 2.3(b), right panel. We can see that the author *G. Parisi* (who is well known for having made significant contributions in different fields of physics) acts as a hub in the community. Different communities can be associated with his different fields of interest.

Now, we compare the algorithm EAGLE with Newman's fast algorithm and the $k$-clique algorithm by applying them to the scientific collaboration network. Figure 2.4 shows that the hierarchical community structure found by Newman's fast



(a) Word association network

(b) Scientific collaboration network

**Fig. 2.3** The hierarchical and overlapping community structure in real world networks. The two networks are (**a**) the word association network, and (**b**) the scientific collaboration network. Each *numbered circle* denotes a community and the *number in the circle* denotes its size. Communities connected by a link overlap with each other. Different communities are rendered in different colors (or markers for print). The overlapping nodes and edges between communities are colored in *red*. In addition, the values of the corresponding *EQ* are also given when breaking networks (communities) down into communities (sub-communities). Reprinted from Ref. [30], Copyright 2009, with permission from Elsevier

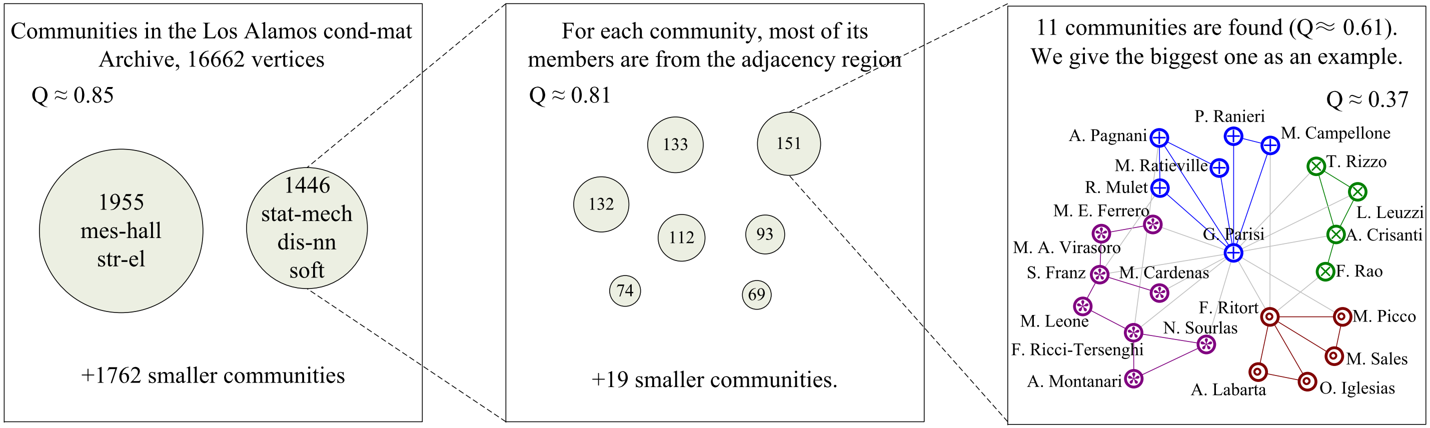


**Fig. 2.4** The hierarchical community structure found by Newman's fast algorithm in the scientific collaboration network. Each *numbered circle* denotes a community and the *number in the circle* denotes its size. Communities connected by a link overlap with each other. Different communities are rendered in different colors (or markers for print). The overlapping nodes and edges between communities are colored in *red*. In addition, the values of the corresponding *Q* are also given when breaking networks (communities) down into communities (sub-communities). Reprinted from Ref. [30], Copyright 2009, with permission from Elsevier

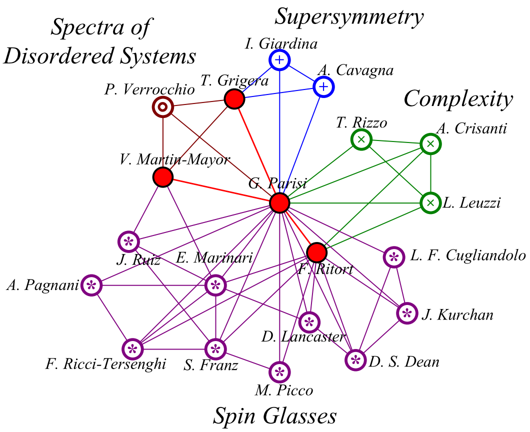


**Fig. 2.5** The overlapping community structure around the node *G. Parisi* in the scientific collaboration network. Different communities are rendered in different colors (or markers for print). The overlapping nodes and edges between communities are colored in *red*. Here, $k$ is set to be 4. Reprinted from Ref. [30], Copyright 2009, with permission from Elsevier

algorithm. The number of communities at each level of the hierarchy is almost identical to that found by the algorithm EAGLE although the size of each community is somewhat different. Compare the left panel of Fig. 2.4 with that of Fig. 2.3(b), one community disappears. Actually, it is divided into several other smaller communities, which are not depicted. As to the right panels, the details of communities were given. The node *G. Parisi*, acting as a hub in Fig. 2.3, only appears in one community in Fig. 2.4. The reason is that Newman's algorithm gives rise to partitions of network, while the algorithm EAGLE allows overlaps between communities. Note that overlap between communities is a very common phenomenon in real networks and may contribute to the evolvement of communities and the dynamics of networks.

Figure 2.5 shows the overlapping community structure around the node G. Parisi in the scientific collaboration network. Compare to Fig. 2.3, both the algorithm EAGLE and the $k$-clique algorithm can find the overlapping community structure, although the overlapped communities found by the two algorithms are somewhat different. However, the algorithm EAGLE can give the hierarchy of these overlapped communities compared to the $k$-clique algorithm. The hierarchy of communities is useful to understand the community structure of real world networks.

## 2.3 Extending Modularity to Quantify the Overlapping Community Structure

In this section, a measure for the quality of a cover is proposed to quantify the overlapping community structure referred as $Q_c$ (quality of a cover). With the measure $Q_c$, the overlapping community structure can be identified by finding an optimal cover, i.e., the one with the maximum $Q_c$. The $Q_c$ is based on a maximal clique view of the original network. A maximal clique is a clique (i.e. a complete subgraph) which is not a subset of any other clique in a graph. The maximal clique view is according to a reasonable assumption that a maximal clique cannot be shared by two communities due to that it is highly connective. To find an optimal cover, we construct a maximal clique network from the original network. We then prove that the optimization of $Q_c$ on the original network is equivalent to the optimization of the modularity on the maximal clique network. Thus the overlapping community structure can be identified through partitioning the maximal clique network with an efficient modularity optimization algorithm, e.g., the fast unfolding algorithm in [34]. The effectiveness of the measure $Q_c$ is demonstrated by extensive tests on both the artificial networks and the real world networks with known community structure and the application to the word association network.

### *2.3.1 Quantifying the Overlapping Community Structure*

Before introducing how to quantify the overlapping community, we first illustrate the representation of overlapping community. Figure 2.6 shows an example network with overlapping community structure. The overlapping community structure can be represented by a cover of network. A cover of network is defined as a set of clusters such that each node is assigned to one or more clusters and no cluster is a proper subset of any other cluster. As to the network in Fig. 2.6, the overlapping community structure can be represented by the cover $\{\{1, 2, 3, 4, 5, 6\}, \{3, 7, 8, 9, 10, 11, 12, 13\}, \{10, 11, 12, 14, 15, 16, 17\}, \{18, 19, 20, 21, 22, 23, 24\}\}$.

We have known that the overlapping community structure can be represented as a cover of network instead of a partition of network. Therefore, the overlapping community structure can be quantified through a measure of a cover of network.

As well known, the modularity was used to measure the goodness of a partition of network. Given an un-weighted, undirected network $G(E, V)$ and a partition $P$ of the network $G$, the modularity can be formalized as

$$Q = \frac{1}{L}\sum_{c\in P}\sum_{vw}\delta_{vc}\delta_{wc}\left(A_{vw} - \frac{k_v k_w}{L}\right), \tag{2.3}$$

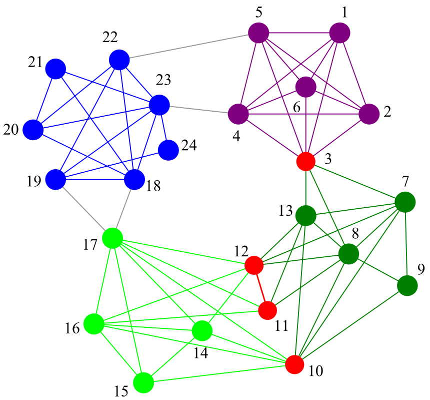


**Fig. 2.6** A schematic network with overlapping community structure. Communities are differentiated by colors and the overlapping regions are emphasized in *red*. The edges between communities are colored in *gray*. Reprinted from Ref. [35], Copyright 2009, with permission from IOP Publishing and SISSA

where $A$ is the adjacency matrix of the network $G$, $L = \sum_{vw} A_{vw}$ is the total weight of all the edges, and $k_v = \sum_w A_{vw}$ is the degree of the node $v$.

In Eq. 2.3, $\delta_{vc}$ denotes whether the node $v$ belongs to the community $c$. The value of $\delta_{vc}$ is 1 when the node $v$ belongs to the community $c$ and 0 otherwise. For a cover of network, however, a node may belong to more than one community. Thus $\delta_{vc}$ needs to be extended to a belonging coefficient $\alpha_{vc}$, which reflects how much the node $v$ belongs to the community $c$.

With the belonging coefficient $\alpha_{vc}$, the goodness of a cover $C$ can be measured by

$$Q_c = \frac{1}{L}\sum_{c\in C}\sum_{vw}\alpha_{vc}\alpha_{wc}\left(A_{vw} - \frac{k_v k_w}{L}\right). \tag{2.4}$$

The idea of the belonging coefficient was proposed in [24]. Its authors also pointed out that the belonging coefficient should satisfy a normalization property. This property is formally written as

$$0 \le \alpha_{vc} \le 1, \quad \forall v \in V,\ \forall c \in C \tag{2.5}$$

and

$$\sum_{c\in C}\alpha_{vc} = 1. \tag{2.6}$$

Equations 2.5 and 2.6 only give the general constraints on $\alpha_{vc}$, which lead to such a huge solution space that the enumeration of all the solutions is impractical. To reduce the solution space and make the problem tractable, we introduce an additivity property for the belonging coefficient: the belonging coefficient of a node to a community $c$ is the sum of the belonging coefficients of the node to all of $c$'s sub-communities.

For example, we assume that $C = \{c_1, c_2, \ldots, c_{r-1}, c_r, \ldots, c_s, c_{s+1}, \ldots, c_n\}$ is a cover of the network $G$ and $C' = \{c_1, c_2, \ldots, c_{r-1}, c_u, c_{s+1}, \ldots, c_n\}$ is another cover of $G$. The difference between $C'$ and $C$ is that the community $c_u$ is the union of the communities $c_r, \ldots, c_s$. The additivity property of belonging coefficient can then be formally denoted as

$$\alpha_{vc_u} = \sum_{i=r}^{s}\alpha_{vc_i}. \tag{2.7}$$

The belonging coefficient $\alpha_{vc}$ reflects how much a node $v$ belongs to a community $c$. Intuitively, it is proportional to the total weight of the edges connecting the node $v$ to the nodes in the community $c$, i.e.,

$$\alpha_{vc} \propto \sum_{w\in V(c)} A_{vw}, \tag{2.8}$$

where $V(c)$ denotes the set of nodes belonging to community $c$. Note that the additivity property of belonging coefficient requires that communities are disjoint from a proper view of the network. Therefore, we introduce the maximal clique view to achieve this purpose. We define $\alpha_{vc}$ as the form

$$\alpha_{vc} = \frac{1}{\alpha_v} \sum_{w \in V(c)} \frac{O^c_{vw}}{O_{vw}} A_{vw}, \tag{2.9}$$

where $O_{vw}$ denotes the number of maximal cliques containing the edge $(v, w)$ in the whole network, $O^c_{vw}$ denotes the number of maximal cliques containing the edge $(v, w)$ in the community $c$, and $\alpha_v$ is a normalization term denoted as

$$\alpha_v = \sum_{c \in C} \sum_{w \in V(c)} \frac{O^c_{vw}}{O_{vw}} A_{vw}. \tag{2.10}$$

Obviously, the definition of $\alpha_{vc}$ in Eq. 2.9 satisfies the normalization property. It also satisfies the additivity property if we assume that each maximal clique only belongs to one community. This assumption is reasonable since a maximal clique is highly connective that any two communities sharing a maximal clique should be combined into a single one.

With Eqs. 2.4 and 2.9, we obtain the detailed form of $Q_c$ as a measure to the quality of a cover of network. Note that when a cover degrades to a partition, $Q_c$ becomes the modularity $Q$ in [15] accordingly. In addition, $Q_c = 0$ when all nodes belong to the same community, and it will be shown later in Sect. 2.3.4 that a high value of $Q_c$ indicates a significant overlapping community structure.

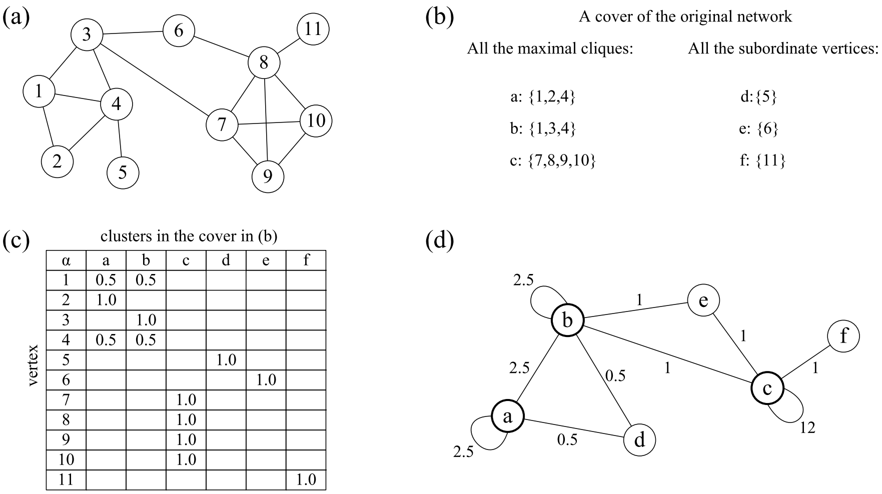
## 2.3.2 *Identifying the Overlapping Community Structure*

With the measure $Q_c$, the overlapping community structure of network can be identified by finding the optimal cover with maximum $Q_c$. To find the optimal cover, we construct a maximal clique network from the original network. Then the overlapping community structure can be identified through partitioning the maximal clique network.

### 2.3.2.1 Construction of the Maximal Clique Network

Given an un-weighted, undirected network $G$, a corresponding maximal clique network $G'$ can be constructed through the following method.

The maximal clique network $G'$ is constructed by defining its nodes and edges. We first find out all the maximal cliques in $G$. We can simply take all these maximal cliques as nodes of $G'$. In practice, however, we observe that some maximal cliques would not be so highly connective if their sizes are too small. Such a maximal clique either lies between different communities (e.g., the maximal cliques $\{4, 23\}$ and $\{5, 22\}$ in the network shown in Fig. 2.6) or connects a node to the whole network



A cover of the original network

All the maximal cliques:

a: {1,2,4}

b: {1,3,4}

c: {7,8,9,10}

All the subordinate vertices:

d:{5}

e: {6}

f: {11}

clusters in the cover in (b)

vertex

| α | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| 1 | 0.5 | 0.5 | | | | |
| 2 | 1.0 | | | | | |
| 3 | | 1.0 | | | | |
| 4 | 0.5 | 0.5 | | | | |
| 5 | | | | 1.0 | | |
| 6 | | | | | 1.0 | |
| 7 | | | 1.0 | | | |
| 8 | | | 1.0 | | | |
| 9 | | | 1.0 | | | |
| 10 | | | 1.0 | | | |
| 11 | | | | | | 1.0 |

**Fig. 2.7** Illustration for the construction process of the maximal clique network. Here, (**a**) The original example network. (**b**) A cover of the original network. In this cover, each maximal clique is a cluster and each subordinate node forms a cluster consisting of only one node. (**c**) The belonging coefficient of each node to its corresponding clusters in the cover. (**d**) The maximal clique network constructed from the example network. Here the parameter $k = 3$. Reprinted from Ref. [35], Copyright 2009, with permission from IOP Publishing and SISSA

(e.g., the maximal clique {8, 11} in the network shown in Fig. 2.7(a)). To deal with these small maximal cliques, we introduce a threshold $k$. Specifically, given the parameter $k$, we only refer to those maximal cliques with the size no smaller than $k$ as the maximal cliques, and refer to those with the size smaller than $k$ as subordinate maximal cliques. We then denote the nodes only belonging to subordinate maximal cliques as subordinate nodes. In this way, each maximal clique or subordinate node in the original network $G$ is taken as one node of $G'$.

Note that all the subordinate nodes and the maximal cliques form a cover $C$ of the original network $G$. For a subordinate node $v$ and a cluster $c$ in the cover $C$, the value of $\alpha_{vc}$ is defined to be 1.0 when $v$ belongs to the cluster $c$ and 0.0 otherwise. As to other nodes, $\alpha_{vc}$ can be obtained according to Eq. 2.9.

Now we can define the edge of the maximal clique network $G'$ by defining its adjacency matrix $B$. Let $m_x$ denote the set of the original network's nodes corresponding to the $x$th node in $G'$. The element of $B$ is defined as

$$B_{xy} = \sum_{vw} \alpha_{vm_x} \alpha_{wm_y} A_{vw} \tag{2.11}$$

and the strength (degree) of the $x$th node

$$s_x = \sum_{y} B_{xy} = \sum_{v} \alpha_{vm_x} k_v. \tag{2.12}$$

For clarity, Fig. 2.7 illustrates the construction process of the maximal clique network from an example network with the parameter $k = 3$. Figure 2.7(b) shows the subordinate nodes and the maximal cliques. Each of them becomes a node in the resulting maximal clique network. For example, the maximal clique $\{1, 2, 4\}$ corresponds to the node $a$ and the subordinate node $\{5\}$ corresponds to the node $d$. Each of these maximal cliques or subordinate nodes is a cluster in a cover $C$ of the original network. Their belonging coefficients corresponding to the cover $C$ are shown in Fig. 2.7(c). According to these belonging coefficients and Eq. 2.11, the weight of each edge of the maximal clique network is obtained. Take the edge connecting the nodes $a$ and $b$ as an example. As known, the node $a$ corresponds to the maximal clique $\{1, 2, 4\}$ and the node $b$ corresponds to the maximal clique $\{1, 3, 4\}$. Using Eq. 2.11, the weight of this edge is $\alpha_{1a}\alpha_{3b} + \alpha_{1a}\alpha_{4b} + \alpha_{2a}\alpha_{1b} + \alpha_{2a}\alpha_{4b} + \alpha_{4a}\alpha_{1b} + \alpha_{4a}\alpha_{3b} = 0.5 + 0.25 + 0.5 + 0.5 + 0.25 + 0.5 = 2.5$.

The constructed maximal clique network is a weighted network though the original network is un-weighted. The total weight $L'$ of all the edges in the maximal clique network is equal to the total weight (number) $L$ of edges in the original network. The proof is

$$\begin{aligned} L' &= \sum_{xy} B_{xy} \\ &= \sum_{xy}\sum_{vw} \alpha_{vm_x}\alpha_{wm_y} A_{vw} \\ &= \sum_{vw} A_{vw} \sum_{x} \alpha_{vm_x} \sum_{y} \alpha_{wm_y} \\ &= \sum_{vw} A_{vw} \\ &= L. \end{aligned} \tag{2.13}$$

Each node in the original network corresponds to more than one node in the maximal clique network. For example, in Fig. 2.7, the node 1 corresponds to two nodes $a$ and $b$ in the maximal clique network. Thus, a partition of the maximal clique network can be mapped to a cover of the original network, which holds the information about the overlapping community structure of the original network.

### 2.3.2.2 Finding the Overlapping Community Structure

Now we investigate the overlapping community structure of the original network through partitioning its corresponding maximal clique network. To find the natural partition of a network, the optimization of modularity is the widely used technique. The partition with the maximum modularity is regarded as the optimal partition of network. We employ the algorithm proposed in [34] to partition our maximal clique network. As an example, Fig. 2.8 shows the partition of a maximal clique network. Different parts of the partition are differentiated by shapes or colors.

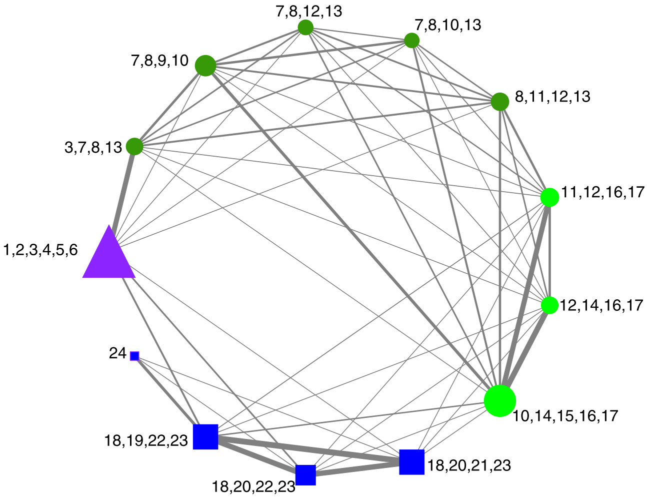


**Fig. 2.8** The maximal clique network constructed from the schematic network in Fig. 2.6. The label near each node shows its corresponding nodes in the original network. The width of line indicates the weight of the corresponding edge. The self-loop edge of each node is omitted and its width is reflected by the volume of the associated circles, squares or triangles. In addition, the optimal partition of the maximal clique network is also depicted. The communities in this partition are differentiated by shapes. Furthermore, the circle-coded community can be partitioned into two sub-communities. The four communities are shown in different colors, which are identical to the communities depicted in Fig. 2.6. Here $k$ is 4. Reprinted from Ref. [35], Copyright 2009, with permission from IOP Publishing and SISSA

As mentioned above, each partition of the maximal clique network corresponds to a cover of the original network and the cover tells us the overlapping community structure. The key problem lies in that whether the optimal partition of the maximal clique network corresponds to the optimal cover of the original network. To answer this question, we analyze the relation between the modularity of the maximal clique network and the $Q_c$ of the original network.

Let $\mathscr{P} = \{p_1, p_2, \ldots, p_l\}$ be a partition of the maximal clique network and $\mathscr{C} = \{c_1, c_2, \ldots, c_l\}$ be the corresponding cover of the original network. Here, $l$ is the size of $\mathscr{P}$ or $\mathscr{C}$, i.e., the number of communities. Using modularity, the quality of the partition $\mathscr{P}$ can be measured by

$$Q = \frac{1}{L'} \sum_i \sum_{x,y \in p_i} \left( B_{xy} - \frac{s_x s_y}{L'} \right). \tag{2.14}$$

Using Eqs. 2.11 and 2.12, we have

$$Q = \frac{1}{L'} \sum_i \sum_{x,y \in p_i} \left( \sum_{vw} \alpha_{vm_x} \alpha_{wm_y} A_{vw} - \frac{1}{L'} \sum_v \alpha_{vm_x} k_v \sum_w \alpha_{wm_y} k_w \right)$$

$$
\begin{aligned}
&= \frac{1}{L'} \sum_i \sum_{x,y \in p_i} \sum_{vw} \alpha_{vm_x} \alpha_{wm_y} \left( A_{vw} - \frac{k_v k_w}{L'} \right) \\
&= \frac{1}{L} \sum_i \sum_{vw} \alpha_{vc_i} \alpha_{wc_i} \left( A_{vw} - \frac{k_v k_w}{L} \right) \\
&= Q_c. 
\end{aligned} \tag{2.15}
$$

Equation 2.15 tells us that the optimization of the $Q_c$ on the original network is equivalent to the optimization of the modularity on the maximal clique network. Thus we can find the optimal cover of the original network by finding the optimal partition of the corresponding maximal clique network. The optimal cover reflects the overlapping community structure of the original network.

### *2.3.3 Discussions*

As to our method, it is important to select an appropriate parameter $k$. On one hand, the parameter $k$ affects the constituent of the overlapping regions between communities. According to the definition to subordinate nodes, they are excluded from the overlapping regions. Thus the larger the parameter $k$, the less the number of nodes which can occur in the overlapping regions. When $k \to \infty$, the maximal clique network is identical to the original network and no overlap is identified. On the other hand, since the subordinate maximal cliques are not so highly connective, the parameter $k$ should not be too small in practice. The choice of the parameter $k$ depends on the specific networks. Observed from many real world networks, the typical value of $k$ is often between 3 and 6. Additionally, as to the networks where larger cliques are rare, our method is close to the traditional modularity-based partition methods. In this case, rare overlaps will be found.

Both the traditional modularity and the $Q_c$ are based on the significance of link density in communities compared to a null-model reference network, e.g., the configuration model network. However, differently from the traditional modularity which requires that each node can only belong to one community, $Q_c$ requires that each maximal clique can only belong to one community. In this way, $Q_c$ takes advantage of both the local topological structure (i.e., the maximal clique) and the global statistical significance of link density.

The same to the traditional modularity, however, the measure $Q_c$ also suffers the resolution limit problem [17], especially when applied to large scale complex networks. Recently, some methods [36] have been proposed to address the resolution limit problem of modularity. These methods are also appropriate to the measure $Q_c$.

Now we turn to the efficiency of our method. It is difficult to give an analytical form of the computational complexity of our method. Here we only discuss what influences the efficiency of our method. Our method consists of three stages, finding out the maximal cliques, constructing the maximal clique network and partitioning the maximal clique network. As to the first stage, we need to find out all the maximal cliques in the network. It is widely believed to be a non-polynomial problem. However, for real world networks, finding all the maximal cliques is easy due to the sparseness of these networks. The computational complexity of the second stage depends on the number of edges in the original networks. Finally, the partition stage rests with the number of the maximal cliques and subordinate nodes. Taken together, our method is very efficient on real world networks.

In addition, as mentioned above, the overlapping community structure can be identified by the optimization of $Q_c$. Similarly, iteratively applying this method to each community, we can investigate the sub-communities correspondingly. In this way, a rigid hierarchical relation of overlapping communities can be identified from the whole network.

## *2.3.4 Results*

In this section, we extensively test our method on the artificial networks and the real world networks with known community structure. Then we apply our method to a large real world complex network, which has been shown to possess overlapping community structure.

### 2.3.4.1 Tests on Artificial Networks

To test our method, we utilize the benchmark proposed in [37]. It provides benchmark networks with heterogeneous distributions of node degree and community size. In addition, it allows for the overlaps between communities. This benchmark poses a much more severe test to community detection algorithms than Newman's standard benchmark [9]. There are many parameters to control the generated networks in this benchmark, the number of nodes $N$, the average node degree $\langle k\rangle$, the maximum node degree $max\_k$, the mixing ratio $\mu$, the exponent of the power-law node degree distribution $t1$, the exponent of the power-law distribution of community size $t2$, the minimum community size $min\_c$, the maximum community size $max\_c$, the number of overlapped nodes $on$, and the number of memberships of each overlapped node $om$. In our tests, we use the default parameter configuration where $N = 1000$, $\langle k\rangle = 15$, $max\_k = 50$, $t1 = 2$, $t2 = 1$, $min\_c = 20$, $max\_c = 50$, $on = 50$ and $om = 2$. By tuning the parameter $\mu$, we test the effectiveness of our method on the networks with different fuzziness of communities. The larger the parameter $\mu$, the fuzzier the community structure of the generated networks is.

To evaluate the effectiveness of an algorithm for the identification of overlapping community structure, a measure is needed to compare the cover found by the algorithm with the ground truth. In [26], a measure is proposed to compare two covers, which is an extension form of *variation of information*. The more similar two covers are, the higher the value of the measure is. Here, we adopt it to compare the overlapping community structure found by our method and the known overlapping community structure in the benchmark networks.

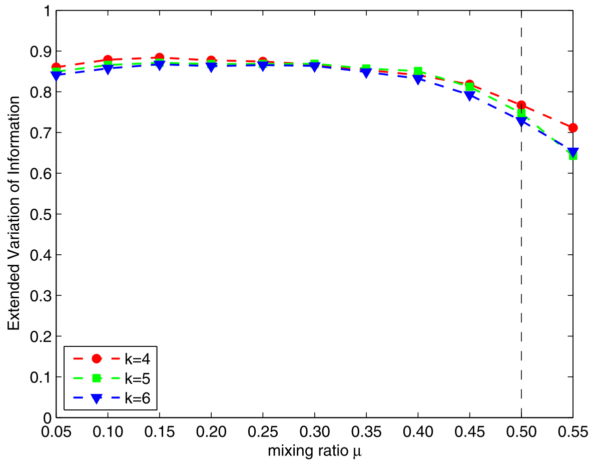


**Fig. 2.9** Test of our method on the benchmark networks. The parameter $k$ in the legend corresponds to the parameter $k$ in our method. The threshold $\mu = 0.5$ (*dashed vertical line* in the figure) marks the border beyond which communities are no longer defined in the strong sense [10], i.e., such that each node has more neighbors in its own community than in the others. Each point corresponds to an average over 100 graph realization. Reprinted from Ref. [35], Copyright 2009, with permission from IOP Publishing and SISSA

Figure 2.9 shows the results of our method with $k = 4, 5, 6$ on the benchmark networks. Our method gives rather good results when the $\mu$ is smaller than 0.5. All of the values of the variation of information are above 0.8. Note that in these cases, communities are defined in the strong sense [10], i.e., each node has more neighbors in its own community than in the others. We also test other settings of $k$ which are larger than 6, and find similar results.

#### 2.3.4.2 Tests on Real World Networks

Our first real world network for test is Zachary's karate club network [38], which is widely used as a benchmark for the methods of community identification. This network characterizes the social interactions between the individuals in a karate club at an American university. A dispute arose between the club's administrator and its principal karate teacher and as a result the club eventually split into two smaller clubs, centered around the administrator and the teacher respectively. The network and its fission is depicted in Fig. 2.10. The administrator and the teacher are represented by nodes 1 and 33 respectively.

Feeding this network into our method with the parameter $k = 4$, we obtain the result shown in Fig. 2.10. Similar to many existing community detection methods, our method partitions the network into four communities. This partition corresponds to the modularity with the value 0.417, while the real partition into two sub-networks has a modularity 0.371. Actually, no node is misclassified by our method. The real

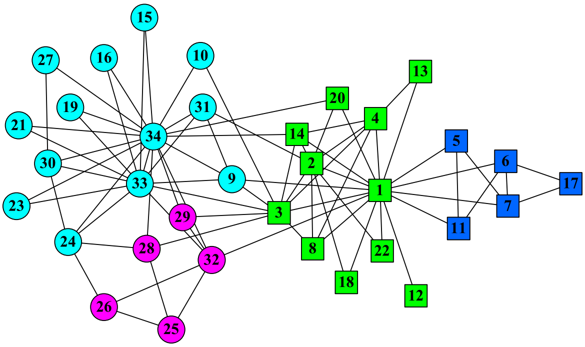


**Fig. 2.10** The network of the karate club studied by Zachary [38]. The real social fission of this network is represented by two different *shapes*, *circle* and *square*. The different colors show the partition obtained by our method with the parameter $k = 4$. Reprinted from Ref. [35], Copyright 2009, with permission from IOP Publishing and SISSA

split of the network can be obtained exactly by pair-wise merge of the four communities found by our method.

We also note that no overlaps are found when $k = 4$. Actually, no overlaps can be found when $k$ is no smaller than 4 as to this network. Overlaps between communities emerge when the parameter $k$ is set to 3. The value of $Q_c$ corresponding to the resulting cover is 0.385 and in total three overlapped communities are found by our method. They are $\{1, 5, 6, 7, 11, 17\}$, $\{1, 2, 3, 4, 8, 9, 12, 13, 14, 18, 20, 22\}$ and $\{3, 9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34\}$. The overlapping regions consist of three nodes, being 1, 3 and 9. Each of them is shared by two communities. Such nodes are often misclassified by traditional partition-based community detection methods. Except the nodes occurring in the overlapping regions, other nodes reflects the real split of the network.

We also test our method on another real world network, a social network of 62 bottlenose dolphins living in Doubtful Sound, New Zealand. The network was constructed by Lusseau [39] with ties between dolphin pairs being established by observation of statistically significant frequent association. The network splits naturally into two groups, represented by the squares and circles in Fig. 2.11.

By applying our method with $k = 4$ to this network, four communities are obtained, denoted by different colors in Fig. 2.11. The green community is connected loosely to the other three ones. Regarding the three circle-denoted communities as a sole community, it and the green community correspond to the known division observed by Lusseau [39]. Furthermore, the three circle-denoted communities also correspond to a real division among these dolphins. The further division appears to have some correlation with the gender of these animals. The blue one consists mainly of females and the other two almost entirely of males.

Alike to the Zarchay's karate network, the overlaps between communities cannot be detected when the parameter $k$ is not less than 4. When $k = 3$, overlaps between the circle-denoted communities emerge while the green community keeps almost intact. The $Q_c$ is 0.490 as to the resulting cover. The nodes occurring in overlapping regions are *Beak*, *Kringel*, *MN*105, *Oscar*, *PL*, *SN*4, *SN*9 and *TR*99 among which the nodes *Beak* and *Kringel* are shared by all the three circle-denoted communities. Again these overlapping nodes are often misclassified by traditional partition-based methods.

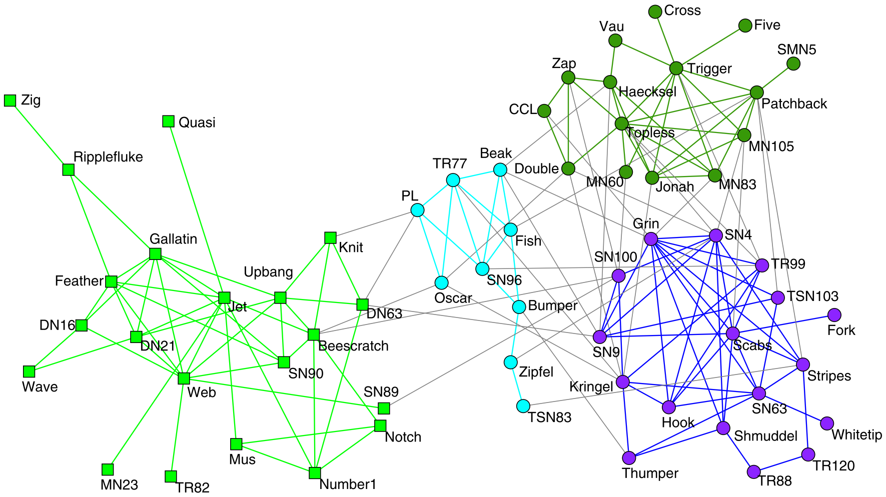


**Fig. 2.11** The community structure of dolphin network. The primary split of the network is represented by different *shapes*, *square* and *circle*. The different colors show the partition obtained by our method with the parameter $k = 4$. Reprinted from Ref. [35], Copyright 2009, with permission from IOP Publishing and SISSA

#### 2.3.4.3 Application to the Word Association Network

Now we apply our method to a large real world complex network, namely the word association network.

The data set for the word association network is from the demo of the software *CFinder* [32]. This network consists of 7207 nodes and 31784 edges, and has been shown to possess overlapping community structure [8]. It is constructed from the South Florida Free Association norms list [40]. Initially, the network is a directed, weighted network. The weight of a directed edge from one word to another indicates the frequency that the people in the survey associated the end point of the edge with its start point. These directed edges were replaced by undirected ones with a weight equal to the sum of the weights of the corresponding two oppositely directed edges. Furthermore, the edges with the weight less than 0.025 were deleted. In this way, an un-weighted, undirected network is obtained, and it is the network we deal with.

Applying our method to the word association network, we obtain in total 20 communities which overlap with each other. The value of the corresponding $Q_c$ is as high as 0.503, indicating a strong overlapping community structure. The size of these found communities are very large that there is no specific semantic meaning for each community. To investigate what is correlated to the overlapping community structure, we apply our method to these communities iteratively and a hierarchy of overlapping communities is obtained. We find that the sub-communities have certain correlation with semantic meaning of words. As an example, Table 2.1 shows us the communities around the word *play*. The five overlapping communities represent different meanings of the word *play*, respectively related to *theater*, *musical instru-*

**Table 2.1** The overlapping communities around the word *play*. Reprinted from Ref. [35], Copyright 2009, with permission from IOP Publishing and SISSA

| No. | Description | Words in each community |
|---|---|---|
| 1 | theater | act actor actress bow character cinema curtsey dance director do drama entertain entertainment film guide involve juggler lead movie participate perform performance *play* portray producer production program scene screen show sing stage television theater |
| 2 | musical instrument | alto band banjo bass beep blues brass bugle cello clarinet clef compose concert conductor country drum faddle fiddle flute guitar harp honk horn instrument jazz keyboard loud music oboe orchestra piano *play* rock saxophone symphony tenor treble trombone trumpet tuba tune viola violin woodwind |
| 3 | children | adults balls children family friends *fun* grown-ups guardians kids love mischief nursery parents *play* playground play_dough prank putty *toy toys* tricycle |
| 4 | sports | active arena athlete athletic baseball basketball black_and_white field football *fun game* illustrated inactive jock pigskin *play* recreation referee soccer sports stadium umpire |
| 5 | toys | board boardwalk checkers chess *fun game* games monopoly nintendo *play* plaything strategy *toy toys* vcr video winning yo-yo |

Note: For each community, a short description is also given. The overlapped words are emphasized in italic type

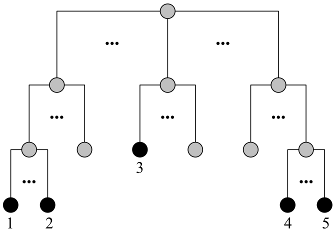


**Fig. 2.12** Part of the hierarchy of communities extracted from the word association network. The *dark-filled circles* correspond to the five communities shown in Table 2.1. Reprinted from Ref. [35], Copyright 2009, with permission from IOP Publishing and SISSA

*ments*, *children*, *sports* and *toys*. Except the common-shared word *play*, four other words are shared by some of these communities. They are *fun*, *game*, *toy* and *toys*. The overlap between these communities characterizes the direct, local relationship between them through sharing members. However, the extent of closeness between communities is sometimes reflected by the indirect, global relationship between them. One of this kind of relationship is the "genealogical" relationship between communities, which can be illustrated by the hierarchy of overlapping communities. Figure 2.12 is an example for hierarchy of communities. As shown in Fig. 2.12, the communities 1 and 2 are in the same branch of the hierarchy, indicating that the meanings represented by them are closer. This can be validated by examining the words contained in these two communities. Similarly, the communities 4 and 5 are also closely related. However, the distance between the communities 3 and 5 is larger although they share as many as 4 words. The overlaps between communities and the hierarchy of these communities provide us a more complete understanding to the relationship between communities.

## 2.4 Conclusions and Discussions

In this chapter, we have studied the problem of detecting both the overlapping and hierarchical community structure in networks. The distinct contribution is that we view a community as consisted of maximal cliques, instead of taking nodes as the building blocks of communities. In this way, the overlapping community structure of networks can be detected under the framework of traditional community detection methods.

Furthermore, representing the overlapping community structure as a cover of network, we propose two kinds of measurements to quantify the quality of a cover of network. The first of is a simple extension of modularity with the consideration that one node can simultaneously belong to more than one community with the same belonging coefficients. For the second one, we proposed a more general extension of modularity (namely $Q_c$) by using a relaxed belonging coefficients. With the $Q_c$ at hand, the overlapping community structure can be detected by optimizing the $Q_c$ to find the optimal cover of network. Then, a maximal clique network is constructed from the original network, and the overlapping community structure can be identified using any modularity optimization method on the maximal clique network.

In addition, $Q_c$ takes advantage of both the local topological structure (i.e., the maximal clique) and the global statistical significance of link density compared with a null-model reference network. In addition, $Q_c$ can be naturally used to simultaneously identify the overlapping and hierarchical community structure of networks. Such a method is helpful to more completely understand the functional and structural properties of networks. The effectiveness of the proposed methods are demonstrated by applications to several real world networks, including the word association network and the scientific collaboration network.

As the further work, we will consider the generalization to the weighted and/or directed networks. It is also an interesting problem about the selection of the parameter $k$ in our method. We will further investigate how to determine an appropriate $k$ for a given network later.

Finally, we give a brief introduction of further readings about overlapping community structure, which has been studied widely in the recent years.

## References

1. Strogatz, S.H.: Exploring complex networks. Nature **410**, 268–276 (2001)
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)
3. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
4. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**, 7821–7826 (2002)
5. Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Nature **433**, 895–900 (2005)
6. Flake, G.W., Lawrence, S.R., Giles, C.L., Coetzee, F.M.: Self-organization and identification of Web communities. IEEE Comput. **35**, 66–71 (2002)

7. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**, 036104 (2006)
8. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)
9. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
10. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA **101**, 2658–2663 (2004)
11. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E **69**, 066133 (2004)
12. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA **103**, 8577–8582 (2006)
13. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**, 036106 (2007)
14. Danon, L., Duch, J., Diaz-Guilera, A., Arenas, A.: Comparing community structure identification. J. Stat. Mech. P09008 (2005)
15. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**, 066111 (2004)
16. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Phys. Rev. E **72**, 027104 (2005)
17. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proc. Natl. Acad. Sci. USA **104**, 36–41 (2007)
18. Kumpula, J.M., Saramaki, J., Kaski, K., Kertesz, J.: Resolution limit in complex network community detection with Potts model approach. Eur. Phys. J. B **56**, 41–45 (2007)
19. Baumes, J., Krishnamoorthy, M., Magdon-Ismail, M., Preston, N.: Finding communities by clustering a graph into overlapping subgraphs. In: Proceedings of IADIS International Conference Applied Computing, pp. 97–104 (2005)
20. Saito, K., Yamada, T., Kazama, K.: Extracting communities from complex networks by the k-dense method. In: Proceedings of the 6th IEEE International Conference on Data Mining, pp. 300–304 (2008)
21. Zhang, S., Wang, R.S., Zhang, X.S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A **374**, 483–490 (2007)
22. Palla, G., Farkas, I.J., Pollner, P., Derényi, I., Vicsek, T.: Directed network modules. New J. Phys. **9**, 186 (2007)
23. Farkas, I.J., Ábel, D., Palla, G., Vicsek, T.: Weighted network modules. New J. Phys. **9**, 180 (2007)
24. Nicosia, V., Mangioni, G., Carchiolo, V., Malgeri, M.: Extending modularity definition for directed graphs with overlapping communities. J. Stat. Mech. P03024 (2009)
25. Evans, T.S., Lambiotte, R.: Line graphs, link partitions, and overlapping communities. Phys. Rev. E **80**, 016105 (2009)
26. Lancichinetti, A., Fortunato, S., Kertesz, J.: Detecting the overlapping and hierarchical community structure of complex networks. New J. Phys. **11**, 033015 (2009)
27. Sales-Pardo, M., Guimerà, R., Moreira, A.A., Amaral, L.A.N.: Extracting the hierarchical organization of complex systems. Proc. Natl. Acad. Sci. USA **104**, 15224–15229 (2007)
28. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.L.: Hierarchical organization of modularity in metabolic networks. Science **297**, 1551–1555 (2002)
29. Pons, P., Latapy, M.: Post-processing hierarchical community structures: Quality improvements and multi-scale view. Theoret. Comput. Sci. **412**, 892–900 (2011)
30. Shen, H.W., Cheng, X.Q., Cai, K., Hu, M.B.: Detect overlapping and hierarchical community structure in networks. Physica A **388**, 1706–1712 (2009)
31. Bron, C., Kerbosch, J.: Finding all cliques in an undirected graph. Commun. ACM 575–577 (1973)
32. Adamcsek, B., Palla, G., Farkas, I.J., Derényi, I., Vicsek, T.: CFinder: Locating cliques and overlapping modules in biological networks. Bioinformatics **22**, 1021–1023 (2006)

33. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. Nature **393**, 440–442 (1998)
34. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. P10008 (2008)
35. Shen, H.W., Cheng, X.Q., Guo, J.F.: Quantifying and identifying the overlapping community structure in networks. J. Stat. Mech. P07042 (2009)
36. Arenas, A., Fernández, A., Gómez, S.: Analysis of the structure of complex networks at different resolution levels. New J. Phys. **10**, 053039 (2008)
37. Lancichinetti, A., Fortunato, S.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E **80**, 016118 (2009)
38. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**, 452–473 (1977)
39. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait? Behav. Ecol. Sociobiol. **54**, 396–405 (2003)
40. Nelson, D.L., McEvoy, C.L., Schreiber, T.A.: The University of South Florida word association, rhyme, and word fragment norms (1998). http://www.usf.edu/FreeAssociation/

# Chapter 3
# Multiscale Community Detection in Networks with Heterogeneous Degree Distributions

## 3.1 Introduction

Graph clustering has been widely applied in exploring regularities emerging in relational data. Recently, the rapid development of network theory correlates graph clustering with the detection of community structure, a common and important topological characteristic of networks. Most existing methods investigate the community structure at a single topological scale. Furthermore, the detection of multiscale community structure is heavily affected by the heterogeneous distribution of node degree. Thus, it is very challenging to detect multiscale community structure in networks with heterogeneous degree distribution.

In the past decade, many methods have been proposed to investigate the community structure in networks [1–3]. These methods identify the community structure through finding an optimal partition of network according to certain criterion or definition of community. In general, the identified community structure corresponds to only one topological scale of network. However, as shown by empirical studies, the community structure of real world networks often exhibits multiple scales [4–7], i.e., more than one topological description is beneficial to characterize the community structure of networks. Thus, it is desired to find methods that can detect multiscale community structure.

Several studies have been conducted to investigate multiple topological scales of networks. Arenas et al. [8] pointed out that synchronization process on networks reveals topological scales of networks and that the spectrum of the Laplacian matrix can be used to identify such topological scales. In Ref. [9], we have used the network conductance to identify multiple topological scales through investigating the diffusion process taking place on networks. Delvenne et al. considered multiscale community structure through investigating the stability of graph communities across time scales [10]. A recent work gave a general optimization framework for the detection of community structure in multiscale networks [11]. Another work considered multiscale community structure through investigating the communities of links instead of communities of nodes [12].

H.-W. Shen, *Community Structure of Complex Networks*, Springer Theses,
DOI 10.1007/978-3-642-31821-4_3, © Springer-Verlag Berlin Heidelberg 2013

However, there are still severe challenges that are not handled by previous work. Existing work does not consider the heterogeneity of node degrees, which is very common to real world complex networks in nature and society. Moreover, it is still an open problem on how to characterize the significance of different relevant topological scales of a complex network.

To address these challenges, we consider the detection of multiscale community structure by introducing a novel framework based on dimensionality reduction. Intuitively, we view the standard representation of network topology as a high-dimensional but redundant description, where each node is taken as one dimension of the network and the edges correspond to data points in the high-dimensional space spanned by these node dimensions. The identification of community structure can be viewed as finding the most significant reduced dimensions that capture the main characteristics of the network topology [13]. Different significance levels for such reduced dimensions correspond to the community structure at different topological scales with different importance at reflecting the characteristics of network topology.

Under the proposed framework, we show that community detection can be viewed as principal component analysis, a major dimensionality reduction method, on the high-dimensional description of networks. Furthermore, we prove that the well-known Laplacian matrix for network partition and the widely-used modularity matrix for community detection are two kinds of covariance matrices used in dimensionality reduction. We then propose a novel method to detect communities at multiple topological scales within our framework. We further show that existing algorithms fail to deal with heterogeneous node degrees. We develop a novel method to handle heterogeneity of networks by introducing a rescaling transformation into the covariance matrices in our framework. Extensive tests on real world and artificial networks demonstrate that the proposed correlation matrices significantly outperform Laplacian and modularity matrices in terms of their ability to identify multiscale community structure in heterogeneous networks.

The remaining of this chapter is organized as follows. Section 3.2 discusses some background. Section 3.3 gives a general framework for the detection of multiscale community structure from the perspective of dimensionality reduction. Section 3.4 presents the rescaling transformation to address the heterogeneity problem. Section 3.5 presents extensive experimental results on a number of artificial and real word networks. Finally, Sect. 3.6 concludes this chapter by highlighting the main contributions and findings.

## 3.2 Preliminaries

### *3.2.1 Principal Component Analysis*

Principal component analysis (PCA) aims to find patterns in data of high dimensions [14]. Here, we give a brief introduction to PCA for the convenience of understanding the remaining part of this chapter.

In PCA, we suppose that we have $m$ data points in an $n$-dimensional space. For convenience, we can represent each data point as an $n$-dimensional column vector and stack all the $m$ column vectors into a data matrix $X$ with the size $n$ by $m$. The empirical mean of these data points is denoted by a column vector $x = \frac{1}{m}\sum_j X_{*j}$, where $X_{*j}$ is the $j$th column of $X$. Furthermore, we make the data matrix $X$ to have zero empirical mean by subtracting the empirical mean $x$ from each column of $X$ and the resulting data matrix is denoted by $\widetilde{X}$. PCA works on the covariance matrix

$$C = \widetilde{X}\widetilde{X}^T/(m-1). \tag{3.1}$$

Here, $m-1$ is used instead of $m$ to make the empirical covariance unbiased when calculated from sample data points rather than a distribution.

PCA transforms high-dimensional data into a small number of principal components, each corresponding to a direction in the space of data points. PCA is defined in such a way that the first principal component accounts for as much of the variance in the data as possible, and each succeeding component in turn has the highest variance under the constraint that it is orthogonal to the preceding components. Without loss of generality, we use a normalized vector $u$ to denote the first principal component. We write $u$ as a linear combination of the normalized eigenvectors $u_i$ of the covariance matrix $C$, i.e., $u = \sum_i^n a_i u_i$, where the coefficients $a_i = u_i^T u$. Since $u$ is a normalized vector, we have $u^T u = 1$ which implies that

$$\sum_{i=1}^{n} a_i^2 = 1. \tag{3.2}$$

The matrix $\widetilde{X}$ can be projected onto the direction $u$ as $u^T\widetilde{X}$. Taking into account that $\widetilde{X}$ has a zero mean, the variance along the direction $u$ can be calculated by

$$\begin{aligned} V &= \frac{1}{m-1}\big(u^T\widetilde{X}\big)\big(u^T\widetilde{X}\big)^T = \frac{1}{m-1}\big(u^T\widetilde{X}\big)\big(\widetilde{X}^T u\big) \\ &= u^T C u = \left(\sum_i^n a_i u_i^T\right) C \left(\sum_j^n a_j u_j\right) \\ &= \sum_{ij} a_i a_j \lambda_j \delta_{ij} = \sum_i a_i^2 \lambda_i, \end{aligned} \tag{3.3}$$

where $\lambda_i$ is the eigenvalue of $C$ corresponding to the eigenvector $u_i$ and we have made use of Eq. 3.1 and $u_i^T u_j = \delta_{ij}$. The function $\delta_{ij}$ is 1 if $i = j$ and 0 otherwise. Without loss of generality, we assume that the eigenvalues are labeled in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The task of maximizing $V$ can then be equated to the task of choosing the nonnegative quantities $a_i^2$ so as to maximize Eq. 3.3 under the constraint in Eq. 3.2.

Obviously, $V$ reaches maximum when we set $a_1^2 = 1$ and $a_i^2 = 0$ $(i \neq 1)$, i.e., the first principal component $u$ is parallel to the eigenvector $u_1$. Then, we turn to the second principal component along which the variance of data points is maximized with the constraint that it is orthogonal to the obtained principal component $u_1$. According to Eqs. 3.2 and 3.3, such a direction is parallel to the eigenvector $u_2$. In a similar way, it can be easily shown that all the eigenvectors of the covariance matrix $C$ are just all the principal components.

Note that $\widetilde{X}$ corresponds to the standard bases $e_i$, whose $i$th element is 1 and other elements are 0. The eigenvectors $u_i$ of the covariance matrix $C$ provide another set of bases. We stack these eigenvectors into a matrix $U$ with its $i$th column being the $i$th eigenvector $u_i$. The data matrix $\widehat{X}$ with respect to the new bases is $\widehat{X} = U^T \widetilde{X}$. Since $\widehat{X}$ is the results of rotating $\widetilde{X}$ around the coordinate origin, $\widehat{X}$ also has zero mean. Using Eq. 3.1, the covariance matrix $\Sigma$ of $\widehat{X}$ can be calculated by

$$\Sigma = \frac{1}{m-1}\widehat{X}\widehat{X}^T = \frac{1}{m-1}U^T \widetilde{X}\widetilde{X}^T U = U^T C U. \tag{3.4}$$

According to Eq. 3.4, $\Sigma$ is a diagonal matrix with its diagonal elements being the eigenvalues of the covariance matrix $C$ in the order corresponding to the eigenvectors stacked in $U$. All the non-diagonal elements of $\Sigma$ are zeroes.

### 3.2.2 Graph Partitioning and the Laplacian Matrix

Graph partitioning problem has a long tradition of research in computer science. This problem is to find a partition of graph with the minimum *cut size*, which is the number of edges between different groups of nodes for unweighted graph. Generally, the desired number of node groups is known a priori. The classical graph partitioning problem deals with two-way partitioning.

An unweighted, undirected graph or network is often described by its adjacency matrix $A$ defined as

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge joining nodes } i, j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

We assume that the network has no self-loop edges. For a two-way partitioning problem, we can use an index vector $s$ to denote the group membership of nodes, i.e., $s_i$ is 1 if node $i$ belongs to the first group and $-1$ otherwise. Then the cut size can be formulated by

$$S = \frac{1}{4}\sum_{ij}(1 - s_i s_j)A_{ij} = \frac{1}{4}\sum_{ij} s_i s_j (k_i \delta_{ij} - A_{ij}), \tag{3.6}$$

where $k_i = \sum_j A_{ij}$ is the degree of node $i$, and $\delta_{ij}$ is 1 if $i = j$ and 0 otherwise. In a matrix form, we have

$$S = \frac{1}{4} s^T L s, \tag{3.7}$$

where $L$ is the Laplacian matrix with its elements

$$L_{ij} = k_i \delta_{ij} - A_{ij} = \begin{cases} k_i & \text{if } i = j, \\ -A_{ij} & \text{otherwise.} \end{cases} \tag{3.8}$$

The matrix $L$ can also be defined in the matrix form $L = D - A$, where $D$ is the diagonal matrix with its $i$th diagonal element being the degree of node $i$.

For the convenience of understanding the matrix $L$ and its later use, we list several properties of $L$.

1. $L$ is symmetric and positive semi-definite.
2. The smallest eigenvalue of $L$ is 0, the corresponding eigenvector is the constant one vector $\mathbb{1}$.
3. $L$ has $n$ non-negative, real-valued eigenvalues $0 = \lambda_1^L \leq \lambda_2^L \leq \cdots \leq \lambda_n^L$.

For proof of these properties and more properties of the matrix $L$, the readers can refer to [15].

According to Eq. 3.7, the cut size $S$ obtains its minimum when the index vector $s$ is parallel to the eigenvector of $L$ corresponding to the smallest eigenvalues. (For details, the readers can refer to [16].) However, such an index vector divides all the nodes into a sole group and this is a trivial solution to the problem of graph partitioning. Thus, of high interest is the eigenvector corresponding to the second smallest eigenvalue, known as the Fiedler's vector. As shown in [17], the Fiedler's vector has been well studied and widely used for two-way graph partitioning. Actually, the Laplacian matrix and its variants play critical role in the spectral theory [18] and have gains success in graph partitioning and image segmentation [19, 20]. More importantly, the Laplacian matrix is often used to characterize the synchronization dynamics on networks. A recent study [8] on the synchronization dynamics on networks showed that the spectrum of the Laplacian matrix reveals the intrinsic topological scales, which are closely related to the community structure of networks.

### 3.2.3 Community Structure and the Modularity Matrix

As a common and important topological characteristic, the community structure is proposed by Girvan and Newman [21]. Since then the community structure has become the research topic of lots of scientific literature [2]. Different from graph partitioning, the detection of community structure aims to find the *natural* partition of networks. Generally, the number of communities and the sizes of communities are not known a priori.

Earlier methods for community detection borrow ideas from traditional hierarchical clustering. They can be roughly classified into agglomerative methods and divisive methods [22–24]. Each of these methods produces a dendrogram and the community structure can be obtained through cutting the dendrogram at an appropriate place according to certain criteria. Each cutting gives rise to a partition of the network. Then, the critical problem becomes choosing the best place to cut the dendrogram.

To address this problem, Newman proposed the modularity as a quality measure for partitions of networks [22]. Given a partition $\mathscr{P}$, the modularity is defined as

$$Q = \frac{1}{2m} \sum_{c \in \mathscr{P}} \sum_{i,j \in c} \left( A_{ij} - \frac{k_i k_j}{2m} \right), \tag{3.9}$$

where $c$ is a community and $2m = \sum_i k_i = \sum_{ij} A_{ij}$ is the total strength of network nodes.

With modularity, the detection of community structure becomes an optimization problem of the modularity among all the possible partitions of a network. Unfortunately, the optimization is proved to be NP-hard [25]. Many heuristic methods are proposed to optimize the modularity, such as greedy algorithms [26, 27], simulated annealing [28], extremal optimization [29], Tabu search [6], and mathematical programming [30].

Recently, in [31], Newman pointed out that the modularity can be expressed in terms of the eigenvectors of a characteristic matrix of the network, which is called the modularity matrix. The elements of the modularity matrix $B$ are written as

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}. \tag{3.10}$$

This matrix gives a spectral explanation of the modularity and provides an effective way to optimize it.

Although the modularity gains great success, it suffers several serious problems which limit its capability and applicability. As pointed out by Fortunato et al. [5], the optimization of modularity faces the resolution limit problem, i.e., the existence of an intrinsic scale beyond which the communities cannot be detected even though these communities are very distinct. Another problem is that only one topological scale is obtained by the optimization of modularity while multiple topological scales exist in real world networks [6]. Finally, as pointed out by us [32], the modularity fails to handle networks with heterogeneous node degrees.

## 3.3 Framework for Detecting Multiscale Community Structure

In this section, we first give a general framework for the detection of multiscale community structure from the perspective of dimensionality reduction. Then, under this framework, we give a unified explanation for the Laplacian matrix for two-way network partition and the modularity matrix for community detection. Finally, using these two matrices as the covariance matrices under our framework, we propose new methods to uncover multiscale community structure of networks.

**Table 3.1** A dimensionality reduction framework for community structure detection. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

| Steps | Descriptions |
|---|---|
| 1 | Give a matrix representation for the topology of network. With this representation, each node corresponds to a vector. |
| 2 | Define the covariance matrix according to the original matrix representation. |
| 3 | Obtain the top $p$ eigenvectors in the descending (ascending) order of eigenvalues if the original matrix representation is positively (negatively) correlated with the intuitive definition of community structure. The value of $p$ is determined according the spectrum of eigenvalues. |
| 4 | Project the original node vectors onto the top eigenvectors and obtain the projected node vectors. |
| 5 | Find the community structure through clustering the projected node vectors. Note that different values of $p$ correspond to community structure at different topological scales. |

## 3.3.1 Our Framework

As described in the previous section, the covariance matrix plays a central role in PCA. Specifically, the eigenvectors of the covariance matrix provide a set of new orthogonal bases to represent the data points. The corresponding eigenvalues characterize the significance of each eigenvector, i.e., principal components in PCA. From the perspective of dimensionality reduction and considering the role of the covariance matrices, we give a general framework for the detection of community structure in networks as shown in Table 3.1.

In our framework, there are two key ingredients. The first one is finding an appropriate covariance matrix of a network for which the PCA analysis on it corresponds to finding community structure. The second one is determining the different topological scales of community structure, i.e., determining the different values of $p$. In the remaining part of this section, we will first show that Laplacian matrix and modularity matrix can be formulated as two kinds of covariance matrices of a network and perform PCA on these covariance matrices to detect community structure. Then, we give the methods to identify multiscale structure using PCA.

## 3.3.2 Covariance Matrices of Networks

We now show that common matrices for community detection, including Laplacian matrix and modularity matrix, can be viewed as covariance matrices in PCA analysis for certain representation of networks.

### 3.3.2.1 Laplacian Matrix

Given a network, we can represent it with a node-edge incidence matrix instead of the adjacency matrix. Specifically, for a directed edge pointing to node $j$ from node $i$, $i$ is called the tail and $j$ is called the head of the edge. The node-edge incidence matrix is defined as

$$Z_{il} = \begin{cases} 1 & \text{if the node } i \text{ is the tail of the edge } l, \\ -1 & \text{if the node } i \text{ is the head of the edge } l, \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

For an undirected edge, it can be replaced by two oppositely directed edges. For clarity, we restrict our attention to undirected networks without self-loop edges although the findings are also applicable to directed networks. We assume that the incidence matrix has the size $n$ by $2m$, where $n$ nodes correspond to $n$ rows and $m$ edges produce the $2m$ columns.

For the incidence matrix $Z$, $n$ nodes correspond to $n$ dimensions. The columns of $Z$ can be taken as $n$-dimensional data points distributed in the space spanned by the $n$ dimensions. Note that the empirical mean of these data points is 0. Thus, the empirical covariance matrix for the $n$ node dimensions can be formulated as

$$C' = \frac{1}{2m-1} ZZ^T, \tag{3.12}$$

with the elements being

$$C'_{ij} = \begin{cases} 2k_i/(2m-1) & \text{if } i = j, \\ -2A_{ij}/(2m-1) & \text{otherwise.} \end{cases} \tag{3.13}$$

Our key observation is that, ignoring the constant $2/(2m-1)$, this covariance matrix $C'$ is identical to the Laplacian matrix $L$ defined in Eq. 3.8. The constant term does not affect the PCA method and it is a common practice to ignore it. In summary, we conclude that the Laplacian matrix is the covariance matrix of network when represented by the node-incidence matrix $Z$.

#### 3.3.2.2 Modularity Matrix

Now we investigate another representation for network, which is given by two node-edge incidence matrices, defined as

$$X_{il} = \begin{cases} 1 & \text{if the node } i \text{ is the tail of the edge } l, \\ 0 & \text{otherwise,} \end{cases} \tag{3.14}$$

and

$$Y_{il} = \begin{cases} 1 & \text{if the node } i \text{ is the head of the edge } l, \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

Note that the rows of $X$ (or $Y$) are mutually orthogonal and that each column sums to unity.

Unlike the matrix $Z$ which represents the direction of edges by the signs of elements, the new representation has two node-edge incidence matrices and the directions of edges are distinguished directly by the different matrices, i.e., the tails of nodes are in the matrix $X$ and the heads of nodes in the matrix $Y$.

For this new representation, the empirical mean of these data points in $X$ is denoted by $x = \frac{1}{2m}\sum_j X_{*j}$, where $X_{*j}$ is the $j$th column of $X$. Similarly, we give the empirical mean of the data points in $Y$ as $y = \frac{1}{2m}\sum_j Y_{*j}$. According to Eqs. 3.14 and 3.15, we have $x = y = (k_1, k_2, \ldots, k_n)^T/2m$. Now we subtract the mean $x$ from each column of $X$ and the mean $y$ from each column of $Y$. Such an operation is known as the *translation* transformation and makes the data points to have a zero mean. The resulting matrices can be denoted by $\widetilde{X} = X - x\mathbb{1}^T$ and $\widetilde{Y} = Y - y\mathbb{1}^T$, where $\mathbb{1}$ is a constant vector with appropriate dimensions. With $\widetilde{X}$ and $\widetilde{Y}$, the empirical covariance between the $i$th row of $X$ and the $j$th row of $Y$ can be calculated by $\widetilde{X}_{i*} \cdot (\widetilde{Y}_{j*})^T/(2m-1)$. As a result, the covariance matrix between all the rows of $X$ and the rows of $Y$ is

$$C'' = \frac{1}{2m-1}\widetilde{X}\widetilde{Y}^T, \tag{3.16}$$

with its elements being

$$C''_{ij} = \frac{1}{2m-1}\left(A_{ij} - \frac{k_i k_j}{2m}\right). \tag{3.17}$$

We can see that, ignoring the constant term $1/(2m-1)$, the covariance matrix $C''$ is identical to the modularity matrix defined in Eq. 3.10. Thus, similar to the Laplacian matrix, the modularity matrix can also be viewed as a kind of covariance matrix of a network. Different from the elements of the Laplacian matrix which are the covariance of the same matrix $Z$, the elements of the modularity matrix is the cross-covariance between the different matrices $X$ and $Y$. This difference will be discussed in the subsequent section when we use these two matrices to study the community structure of networks.

In addition, the derivation of the two covariance matrices can be easily extended to weighted networks if we consider each weighted edge between two nodes as multiple unweighted edges connecting them.

### *3.3.3 Detection of Community Structure as PCA*

We have shown that the Laplacian and modularity matrices are two kinds of covariance matrices of networks. Now we give algorithms for community detection from the perspective of dimensionality reduction. In particular, we show that PCA analysis on these covariance matrices can not only detect community structure, but also identify communities on multiple topological scales.

#### 3.3.3.1 Laplacian Matrix

For a Laplacian matrix, based on our analysis, the off-diagonal elements characterize the covariance between different dimensions corresponding to nodes of a network. If two nodes are connected, the covariance is negative. For networks with community structure, the tail nodes of edges are expected to be positively correlated with the head nodes of edges [34]. Thus, to uncover the community structure of networks, the eigenvectors of a Laplacian matrix should be ranked in an ascending order of the corresponding eigenvalues.

However, the eigenvector corresponding to the smallest eigenvalue of the Laplacian matrix results in a trivial partition of the network, in which all the nodes belong to the same community. According to the properties of the Laplacian matrix, all its eigenvalues are non-negative and only the smallest eigenvalue is 0 for connected networks. Thus we only take into account the eigenvectors corresponding to positive eigenvalues for the purpose of community detection. Note that the Fiedler's vector for network partitioning is the eigenvector corresponding to the smallest positive eigenvalue. As an example, Table 3.2 gives the Fiedler's vector for Zachary's karate club network, which is widely used as a benchmark for community detection. The network and its community structure are depicted in Fig. 3.1. We can see that the real fission of the club network is revealed by the signs of the components in the Fiedler's vector. Only node 3 is misclassified.

**Table 3.2** Eigenvectors corresponding to the least/most positive eigenvalue of the Laplacian/modularity matrix associated with the Zachary's club network. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

| Node Id | Laplacian matrix | Modularity matrix | Node Id | Laplacian matrix | Modularity matrix |
|---|---|---|---|---|---|
| 1 | −0.1121 | −0.3875 | 18 | −0.1002 | −0.1320 |
| 2 | −0.0413 | −0.2696 | 19 | 0.1628 | 0.1394 |
| 3 | 0.0232 | −0.1319 | 20 | −0.0136 | −0.0576 |
| 4 | −0.0555 | −0.2535 | 21 | 0.1628 | 0.1394 |
| 5 | −0.2846 | −0.1340 | 22 | −0.1002 | −0.1320 |
| 6 | −0.3237 | −0.1457 | 23 | 0.1628 | 0.1394 |
| 7 | −0.3237 | −0.1457 | 24 | 0.1557 | 0.2167 |
| 8 | −0.0526 | −0.2094 | 25 | 0.1530 | 0.0563 |
| 9 | 0.0516 | 0.0545 | 26 | 0.1610 | 0.0754 |
| 10 | 0.0928 | 0.0479 | 27 | 0.1871 | 0.1158 |
| 11 | −0.2846 | −0.1340 | 28 | 0.1277 | 0.1028 |
| 12 | −0.2110 | −0.0778 | 29 | 0.0952 | 0.0683 |
| 13 | −0.1095 | −0.1287 | 30 | 0.1677 | 0.2063 |
| 14 | −0.0147 | −0.1350 | 31 | 0.0735 | 0.0963 |
| 15 | 0.1628 | 0.1394 | 32 | 0.0988 | 0.1019 |
| 16 | 0.1628 | 0.1394 | 33 | 0.1303 | 0.3239 |
| 17 | −0.4228 | −0.0585 | 34 | 0.1189 | 0.3698 |

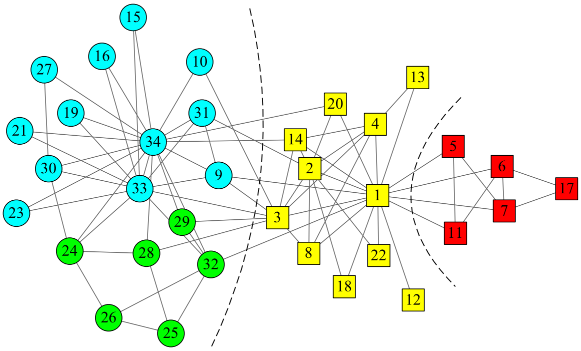


**Fig. 3.1** The network of the karate club network. This network was firstly studied by Zachary [35]. The real social fission of this network is represented by two different shapes, circle and square. This network also exhibits community structure at the other two scales. Specifically, different colors depict the communities at the other scale. The *dashed-curve* gives another alternative partition of network. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

#### 3.3.3.2 Modularity Matrix

For a modularity matrix, the covariance between two node dimensions is positive if the two nodes are connected and negative otherwise. This is consistent with the intuition on community structure, i.e., for networks with community structure, the tail nodes of edges are expected to be positively correlated with the head nodes of edges [34]. Thus, under our framework for community detection, the eigenvectors of the modularity matrix are ranked in descending order of the corresponding eigenvalues and the top eigenvectors provide meaningful information for community structure. As an example, Table 3.2 gives the eigenvector corresponding to the largest eigenvalue of the modularity matrix associated with the Zachary's karate club network. The signs of the components in the eigenvector exactly uncover the real split of the network.

However, different from the Laplacian matrix which is calculated according to the same data matrix $Z$, the modularity matrix is the cross-covariance according to two different data matrices $X$ and $Y$. Thus, the eigenvalues of the modularity matrix can be positive or negative rather than all being non-negative. Generally speaking, the positive eigenvalues indicate that the corresponding eigenvectors make positive contribution to reflect the community structure. As to the negative eigenvalues, the corresponding eigenvectors reflect the so-called anti-community structure where the edges lying among different communities are denser than the edges within communities. When all the eigenvalues are negative, no community structure exists in the network [31]. Thus, for the purpose of community detection, we only consider the eigenvectors corresponding to positive eigenvalues.

In addition, the eigenvectors of the Laplacian and modularity matrices behave very differently. Intuitively, the eigenvectors of the Laplacian matrix characterize the deviation of nodes relative to the center of the networks. As shown in Table 3.2, the Fiedler's vector partitions the network nodes into two groups, denoted by squares and circles in Fig. 3.1. Among the square nodes, node 17 achieves the most negative components in the Fiedler's vector. This indicates that the node 17 is far away from the center of network. Other larger negative components in Fiedler's vector correspond to nodes 6, 7, 5, 11 and 12, which are also very far away from the center of network. Similarly, among the circle nodes, nodes corresponding to larger positive components are also far away from the center of network, such as nodes 27, 30, 15, 16, 19, 21, 23, and 26. For the eigenvectors of the modularity matrix, the magnitude of its components reflects the connectivity of nodes in their respective communities. As shown in Table 3.2, the larger components in the eigenvector of modularity matrix correspond to nodes 1, 2, 4, 33 and 34, which are nodes with high connectivity and central in their communities.

### 3.3.4 Detection of Multiscale Community Structure

An important strength of our framework is that, it provides a natural and effective way to identify communities at multiple scales. As we show before, existing approaches correspond to PCA analysis that only considers the largest positive eigenvalue. Our key observation is that *it is by extending the consideration from the largest to the top few eigenvalues, we can detect multiscale community structure*. This is a unique advantage of our new framework.

To utilize such information provided by all these eigenvectors, we propose a new algorithm that employs the $k$-means clustering method to cluster the node vectors which are obtained through projecting the original coordinate vector of nodes onto the new set of bases $U$, composed of the top eigenvectors. Specifically, with respect to the standard bases, the coordinate vector of the $i$th node is $e_i$. With respect to the new orthogonal bases $U$, the projected node vector becomes $U^T e_i$. Mathematically, the $i$th projected node vector $v_i$ can be denoted by

$$[v_i]_j = U_{ij}. \tag{3.18}$$

We know that direction and magnitude are two critical factors of a vector. As to a node vector, the direction determines the membership of nodes and the magnitude characterizes the degree to which a node belongs to a community. As shown in Table 3.2, the membership of nodes can be determined according to the signs of the components in the eigenvectors, i.e., the direction of the one-dimensional node vectors due to that only one eigenvector is considered. Considering more than one eigenvector can provide more information. Taking the modularity matrix as example, Fig. 3.2 illustrates the role of the direction of two-dimensional node vectors at determining the membership of nodes through considering the eigenvectors corresponding to the two largest eigenvalues. In summary, when we only focus on the assignment of nodes to communities, we can use the normalized node vectors $v_i$ and ignore the magnitude of node vectors.

As a kind of mesoscopic structure of network, the community structure provides a coarse-grained description of the network topology. Hence, only the most signif-

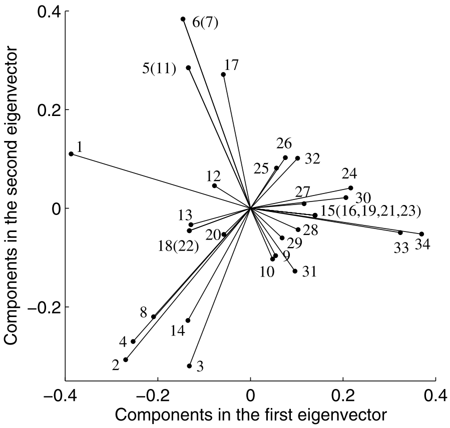


**Fig. 3.2** A plot of the node vectors of the karate club network. Here, the top $p = 2$ eigenvectors of the modularity matrix are utilized. Each node vector is marked by the node index. Several node vectors are identical and thus marked with more than one node indices. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

icant structural features are maintained and the less ones are neglected. Now the tricky problem is how to determine the significant ones. Under our dimensionality reduction framework, this problem is equivalent to choosing the top $p$ significant eigenvectors. We know that the eigenvalues of the covariance matrices characterize the significance of each eigenvector. Hence, we can decide the number of significant eigenvectors based on the corresponding eigenvalues. Intuitively, it is appropriate to choose eigenvectors corresponding to smaller positive eigenvalues of the Laplacian matrix or larger positive eigenvalues of the modularity matrix. Furthermore, a large *eigengap*, i.e., interval between two successive eigenvalues, provides an effective indicator to determine the appropriate number of significant eigenvectors. For the Laplacian matrix, the length of the $i$th eigengap is defined as $\log \lambda_{i+1}^L - \log \lambda_i^L$ $(2 \leq i \leq n-1)$ [8]. As to the modularity matrix, the length of the $i$th eigengap is defined as $\lambda_{i-1}^B - \lambda_i^B$ $(2 \leq i \leq n)$ [32]. Similar methods have been adopted in other contexts to take the advantage of the eigengap of many other types of matrices [8, 9, 16, 36, 37]. The choice of eigenvectors with different significance levels corresponds to the community structure at different topological scales. Our key observation is that *the existence of a significant scale is indicated by the occurrence of a large eigengap*.

Another important problem is the determination of the number of communities. After choosing the $p$ significant eigenvectors, according to Eq. 3.18, each node in the network is represented by a $p$-dimensional node vector through projecting its standard coordinate vector onto the $p$ significant eigenvectors. Then, the identification of community structure amounts to partitioning the node vectors into groups. According to [16], $p + 1$ is the number of communities when the top $p$ eigenvectors are employed to obtain the projected node vectors.

#### 3.3.4.1 Two Examples

Taking the Zachary's karate club network as example again, we illustrate the effectiveness of the eigengap at determining the number of communities. As shown in

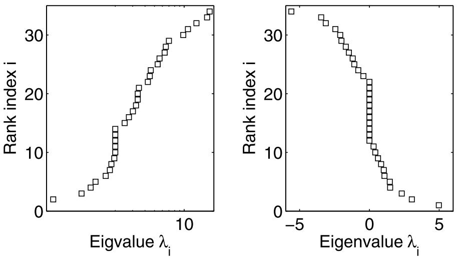


**Fig. 3.3** The spectrum of the covariance matrices associated with the karate club network. *Left panel*: the Laplacian matrix. *Right panel*: the modularity matrix. For the Laplacian matrix, the trivial eigenvalue 0 is ignored and the x-axis is in logarithmic scale. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

Fig. 3.3 (left panel), the largest eigengap of the Laplacian matrix occurs between the first and second smallest positive eigenvalues. This indicates that only the Fiedler's vector is the significant eigenvector and it partitions the network into two communities. The right panel of Fig. 3.3 shows that the largest eigengap of the modularity matrix resides between the largest and the second largest eigenvalues (only the ones between positive eigenvalues are considered for the community detection). It indicates that it is appropriate to utilize only the first eigenvector and the number of communities is 2. The resulting two communities exactly reflect the real split of the network in Fig. 3.1. In addition, for the modularity matrix, besides the largest eigengap, two other relatively larger eigengaps can be observed, one between the second and third largest eigenvalues, and the other between the third and forth. The resulting partition according to these two eigengaps are also depicted in Fig. 3.1, one dividing the network into three communities separated using dashed curves, and the other dividing the network into four communities differentiated by colors. These two partitions are often the results of many traditional community detection methods for a single topological scale [2]. Although they are not identical to the real split of the network, they reveal certain relevant topological feature of the network at alternative scales.

Actually, for a network with multiscale community structure, each scale corresponds to a large eigengap in the spectrum of the covariance matrices. Thus, we can identify the multiscale community structure using the top eigenvectors indicated by these different eigengaps.

As another example, we illustrate the identification of multiscale community structure of the H13-4 network (shown in Fig. 3.5a), which is constructed according to [8]. The network has two predefined hierarchical levels. The first hierarchical level consists of 4 groups of 64 nodes and the second level consists of 16 groups of 16 nodes. On average, each node has 13 edges connecting to the nodes in the same group at the second level and has 4 edges connecting to the nodes in the same group at the first level. This explains the name of such kind of networks. In addition, the average degree of each node is 18. According to the construction rules of the H13-4 network, the two hierarchical levels constitute the different topological description of the community structure of the H13-4 network at different scales.

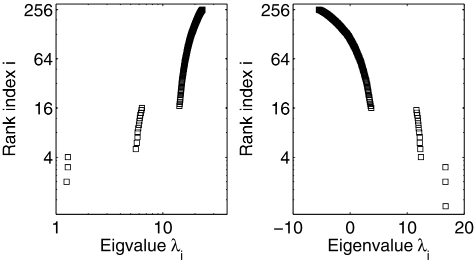


**Fig. 3.4** The spectrum of the covariance matrices associated with the H13-4 network. *Left panel*: the Laplacian matrix. *Right panel*: the modularity matrix. For the Laplacian matrix, the trivial eigenvalue 0 is ignored and the x-axis is in logarithmic scale. The H13-4 network is shown in Fig. 3.5a. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

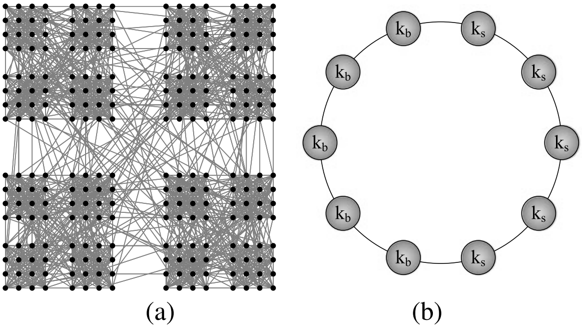


**Fig. 3.5** Two schematic networks. (**a**) The H13-4 network. (**b**) The clique circle network. Each *circle* corresponds to a clique, whose size is marked by its label. The cliques labeled with $k_s$ are smaller cliques with the size $s$, while the cliques labeled with $k_b$ are bigger cliques with the size $b$. Here, $s = 10$ and $b = 20$. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

As shown in Fig. 3.4, two significant eigengaps can be observed in the spectrum of either the Laplacian matrix or the modularity matrix.

- One eigengap occurs between the 3rd and 4th smallest positive eigenvalues for the Laplacian matrix or between the 3rd and 4th largest eigenvalues for the modularity matrix. The topological scale indicated by such eigengaps corresponds to the partition dividing the nodes into 4 groups. Actually, the resulting communities are exactly the predefined 4 groups of 64 nodes in the first hierarchical level.
- The other eigengap occurs between the 15th and the 16th smallest positive eigenvalues for the Laplacian matrix or between the 15th and 16th largest eigenvalues for the modularity matrix. This eigengap indicates the other significant topological scale corresponds to the partition dividing the network nodes into 16 groups. Again, the resulting communities are exactly the predefined 16 groups in the second hierarchical level.

- Although the two intrinsic scales can be identified by either the Laplacian matrix or the modularity matrix, according to the lengths of eigengaps, the Laplacian matrix prefers the partition dividing the network into 4 groups of nodes while the modularity matrix tends to give the partition dividing the network into 16 groups of nodes.

## 3.4 Heterogeneity Problem and the Rescaling Transformation

In the previous section, we have introduced the method to identify the multiscale community structure using a dimensionality reduction framework with the Laplacian and modularity matrices as covariance matrices. In addition, we proposed to use different eigengaps in the spectrum of the covariance matrices to identify community structure at different topological scales. However, this method works well only for homogeneous networks, where the nodes have approximately the same degree and the communities at a specific scale are of the same size. However, real world networks usually have heterogeneous node degrees and community sizes.

We first illustrate the ineffectiveness of the covariance matrices to deal with the heterogeneous node degrees and community sizes using a schematic network, which is often called the clique circle network as depicted in Fig. 3.5b. Generally speaking, the intrinsic community structure corresponds to the partition where each clique is taken as a community, that is, only one intrinsic scale exists in this network. However, as shown in Fig. 3.6 (left panel), two scales are observed when we investigate the community structure of this network using the spectrum of the Laplacian matrix. One scale corresponds to the intrinsic scale of the network, and the other corresponds to dividing the network nodes into 3 groups, which is not desired. Similarly, as shown in Fig. 3.7 (left panel), besides the intrinsic scale of the clique circle network, another undesired scale is observed in the spectrum of the modularity matrix which corresponds to dividing the network nodes into 5 groups. These results demonstrate that the covariance matrices have difficulty in handling heterogeneity of networks.

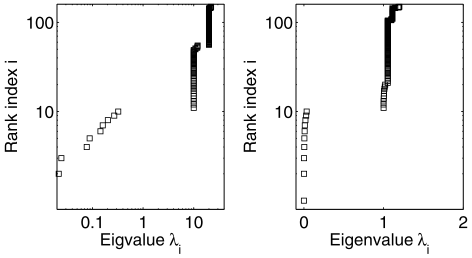


**Fig. 3.6** The spectrum of the Laplacian matrices corresponding to the clique circle network. *Left panel*: the Laplacian matrix. *Right panel*: the normalized Laplacian matrix. The clique circle network is depicted in Fig. 3.5b. The *horizontal axis* shows the eigenvalues of matrix and the *vertical axis* represents the rank index of eigenvalues. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

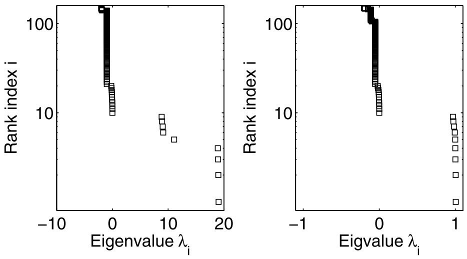


**Fig. 3.7** The spectrum of the modularity matrices corresponding to the clique circle network. *Left panel*: the modularity matrix. *Right panel*: the normalized modularity matrix. The clique circle network is depicted in Fig. 3.5b. The *horizontal axis* shows the eigenvalues of matrix and the *vertical axis* represents the rank index of eigenvalues. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

To address this problem, we reconsider the formulation of these two covariance matrices. When formulating the covariance matrices from the data matrices, defined in Eqs. 3.11, 3.14 and 3.15, the covariance matrices are both zero-centering through subtracting off the mean of data points. This is called the translation transformation. When using the eigenvectors of the Laplacian or modularity matrix as the new orthogonal bases instead of the standard bases, the rotation transformation is utilized. These two transformations, however, do not take into account the difference among the variances of the original dimensions, each corresponding to one node. Thus, the spectrum of the Laplacian and modularity matrices fails to deal with the heterogeneity of node degrees. We propose a remedy called *rescaling transformation*. Through introducing the rescaling transformation into the Laplacian matrix, we obtain the normalized Laplacian matrix, which can be formulated as

$$L^{norm} = (\Sigma_Z)^{-1/2} C' (\Sigma_Z)^{-1/2}, \tag{3.19}$$

where the element $(\Sigma_Z)_{ii} = 2k_i/(2m-1)$ of the diagonal matrix $\Sigma_Z$ denotes the empirical variance of the original data matrix $Z$ defined in Eq. 3.11 along the $i$th axis direction. Specifically, the elements of $L^{norm}$ can be written as

$$L_{ij}^{norm} = \frac{L_{ij}}{\sqrt{k_i k_j}}. \tag{3.20}$$

Similarly, using the rescaling transformation, we obtain the normalized modularity matrix

$$R = (\Sigma_X)^{-1/2} C (\Sigma_Y)^{-1/2}, \tag{3.21}$$

where $\Sigma_X$ is a diagonal matrix with its diagonal elements $(\Sigma_X)_{ii} = k_i(1 - k_i/2m)/(2m-1)$ being the empirical variance of $X$ along the $i$th standard axis, and $\Sigma_Y$ is a diagonal matrix with its diagonal elements $(\Sigma_Y)_{jj} = k_j(1-k_j/2m)/(2m-1)$ being the empirical variance of $Y$ along the $j$th standard axis. Specifically, the elements of $R$ are defined as

$$R_{ij} = \frac{A_{ij} - \frac{k_i k_j}{2m}}{\sqrt{k_i(1 - k_i/2m)}\sqrt{k_j(1 - k_j/2m)}}. \tag{3.22}$$

In statistics, these normalized matrices are called *correlation matrices*.

Compared with the covariance matrices, the correlation matrices have two advantages. Firstly, the correlation matrices can well deal with the heterogeneity of node degrees. As shown in Fig. 3.6 (right panel) and Fig. 3.7 (right panel), the intrinsic scale of the clique circle network is correctly revealed by the spectrum of both correlation matrices. Furthermore, different from the covariance matrices, no undesired topological scales are observed. Secondly, for the correlation matrices, the magnitude of their eigenvalues themselves can provide vital information for the cohesiveness within each community and the looseness of connections between different communities. As shown in Fig. 3.6 (right panel), the eigenvalues on the smaller side of the largest eigengap all approach 0. Similarly, as shown in Fig. 3.7 (right panel), the eigenvalues on the greater side of the largest eigengap all approach 1. Both indicate that the intrinsic communities are very cohesive. Meanwhile, other eigenvalues are very small, indicating connections between different communities are loose.

The second advantage of the correlation matrices is especially important for community detection on networks without a significant topological scale. For these networks, the eigengaps of the covariance matrices or the correlation matrices both fail to provide obvious evidence for the number of intrinsic communities. In these cases, the eigenvalues themselves can provide critical information of the community structure.

In addition, for the original covariance matrices, the magnitude of their eigenvalues is influenced by the network size and the heterogeneity of node degrees. Hence, the eigenvalues cannot provide useful information to determine the cohesiveness of the communities and the number of intrinsic communities. For the correlation matrices with rescaling transformation, however, the magnitude of the eigenvalues has been rescaled and thus can provide information about the cohesiveness of communities and help us choose the desired scale with respect to specific application demands. Moreover, the eigenvalues of the correlation matrices can be compared across different networks since they are rescaled and not influenced by the network size.

## 3.5 Experimental Results

In this section, we empirically demonstrate the effectiveness of the multiscale community detection methods based on our dimensionality reduction framework. In addition, we apply our approach to a variety of real world networks.

### 3.5.1 Tests on Synthetic Benchmark Networks

We utilize the benchmark proposed by Lancichinetti et al. in [38]. This benchmark provides networks with heterogeneous node degrees and community sizes, which are common characteristics in real world networks. Many parameters are used to control the generated networks: the number of nodes $N$, the average node degree $\langle k\rangle$, the maximum node degree max_$k$, the mixing ratio $\mu$, the exponent $\gamma$ of the power law distribution of node degree, the exponent $\beta$ of the power law distribution of community size, the minimum community size min_$c$, and the maximum community size max_$c$. In our tests, we use the default parameter configuration where $N = 1000$, $\langle k\rangle = 15$, max_$k = 50$, min_$c = 20$, and max_$c = 50$. To test the influence of the distribution of node degree and community size, we adopt four parameter configurations for $\gamma$ and $\beta$, $(\gamma, \beta) = (2, 1)$, $(\gamma, \beta) = (2, 2)$, $(\gamma, \beta) = (3, 1)$ and $(\gamma, \beta) = (3, 2)$. By tuning $\mu$, we test the effectiveness of our method on networks with different fuzziness of communities. A larger $\mu$ gives a fuzzier community structure. In addition, we adopt the normalized mutual information (NMI) [1] to compare the partition found by community detection methods against the true partition. A larger NMI indicates a better method.

Note that each benchmark network has only one intrinsic topological scale according to the construction rules. Therefore, we only consider the largest eigengap in the spectrum of the covariance and correlation matrices. The communities are identified using the top $p$ significant eigenvectors indicated by the largest eigengap. The $p$ eigenvectors are projected into the node vectors according to Eq. 3.18, and the communities are identified by clustering these node vectors using the $k$-means clustering method. Note that this results in $p + 1$ communities.

The first test focuses on whether the intrinsic scale can be correctly uncovered. Figure 3.8 shows the comparison between the Laplacian matrix and the normalized Laplacian matrix. Figure 3.9 compares the modularity matrix and the normalized modularity matrix. When the community structure is evident, i.e., the mixing ratio $\mu$ is smaller, both the two covariance matrices and the two correlation matrices are effective at identifying the correct number of communities and thus the intrinsic scale of the network. However, when the community structure becomes fuzzier with an increased $\mu$, the performance of the original covariance matrices deteriorates while the correlation matrices with the rescaling transformation still achieve good results. Even when the mixing ratio $\mu$ is larger than 0.5, the border beyond which communities are no longer defined in the strong sense [23], the number of communities can still be accurately identified by investigating the spectrum of the correlation matrices.

The second test turns to whether the intrinsic community structure can be identified. As demonstrated by the first test, the correlation matrices outperform the covariance matrices at finding the correct number of communities. In the second test, we assume that the community number has been given a priori and then we compare the effectiveness of the eigenvectors of these two kinds of matrices in terms of the NMI. As shown in Figs. 3.10 and 3.11, all the four matrices exhibit very good performance at identifying the intrinsic community structure when the community

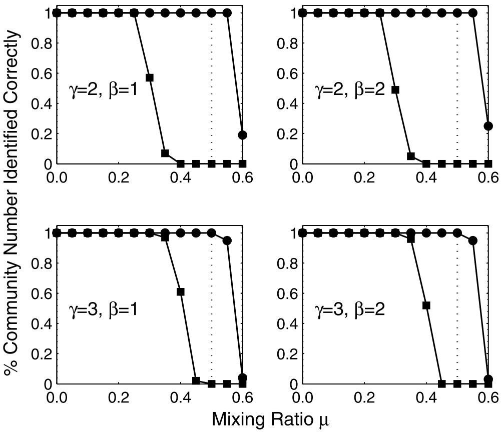


**Fig. 3.8** Comparison between Laplacian matrices at identifying community number. The standard Laplacian matrix is depicted by □ and the normalized Laplacian matrix is depicted by ○. The comparison is conducted on benchmark networks with different parameter configurations. For each parameter configuration, 100 generated networks are used. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

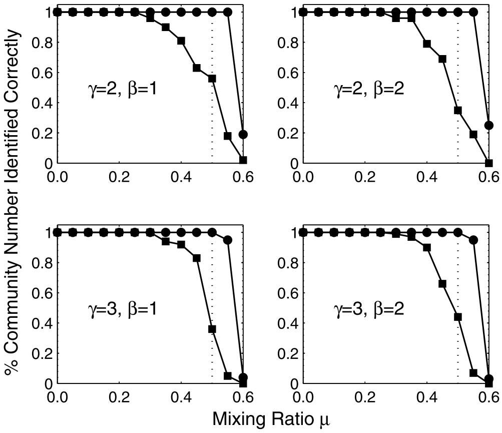


**Fig. 3.9** Comparison between modularity matrices at identifying community number. The modularity matrix is depicted by □ and the normalized modularity matrix is depicted by ○. The comparison is conducted on benchmark networks with different parameter configurations. For each parameter configuration, 100 generated networks are used. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

structure is evident. When the structure is less evident ($\mu$ is larger), however, the correlation matrices outperform the covariance matrices for all parameter configurations. This indicates that the eigenvectors of the correlation matrices characterize the spread characteristics of network nodes better than covariance matrices, especially when the community structure is fuzzier.

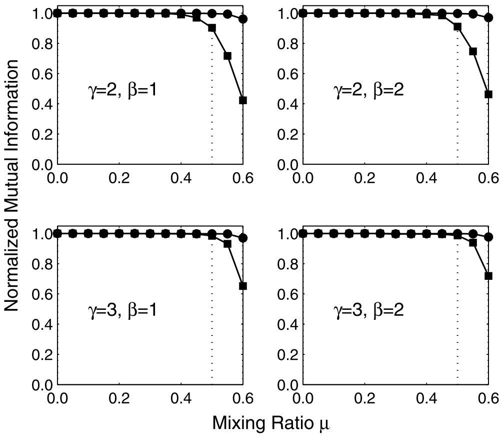


**Fig. 3.10** Comparison between Laplacian matrices at identifying community structure. The standard Laplacian matrix is depicted by □ and the normalized Laplacian matrix is depicted by ○. The comparison is conducted on benchmark networks with different parameter configurations. Each point corresponds to an average over 100 network realizations. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

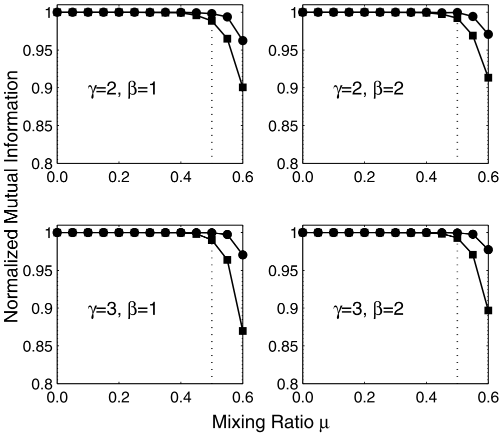


**Fig. 3.11** Comparison between Laplacian matrices at identifying community structure. The modularity matrix is depicted by □ and the normalized modularity matrix is depicted by ○. The comparison is conducted on benchmark networks with different parameter configurations. Each point corresponds to an average over 100 network realizations. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

In addition, the test results indicate that the two correlation matrices behave almost identically although their corresponding two covariance matrices have rather different performance. This phenomenon further indicates that the rescaling transformation is critical for handling heterogeneity in networks.

## 3.5.2 *Tests on Real World Networks*

### 3.5.2.1 Co-author Network: A Case Study

Many real world networks can be used to test the effectiveness of our approach. Here, taking a co-author network as an example, we illustrate the superiority of the correlation matrices to the covariance matrices and later we will only focus on the applications of the correlation matrices on more real world networks.

The nodes of the co-author network represent all individuals who are authors of papers cited in the bibliographies of either of two recent reviews on network research [39, 40] and edges join every pair of individuals whose names appear together as authors of a paper in those bibliographies. The network is constructed as described in [16]. In total, 1589 authors contributed to the papers in the bibliographies and thus the obtained network contains 1589 nodes. We only focus on the giant component of this network, containing 379 nodes and 914 weighted edges.

Investigating the community structure of the network from the perspective of dimensionality reduction, Fig. 3.12 illustrates the spectrum of the Laplacian matrix, the normalized Laplacian matrix, the modularity matrix and the normalized modularity matrix. As to the Laplacian matrix and the modularity matrix, the largest eigengap indicates that the significant topological scale corresponds to the partition dividing the network into 2 groups. As to the two correlation matrices, as shown in the insets in Fig. 3.12(b) and (d), the most significant topological scale corresponds to the partition dividing the nodes into 46 groups. To facilitate the comparison between these two partitions, we gives the topology of the co-author network in Fig. 3.13. Meanwhile, the partition with 46 groups is also depicted with different colors. Intuitively, we can see that such a partition captures the main topological characteristics of the co-author network and provides a much better representation of the community structure than the coarse-grained two-way partition detected by the covariance matrices. This can be further verified through checking the name of each author and his/her research interest.

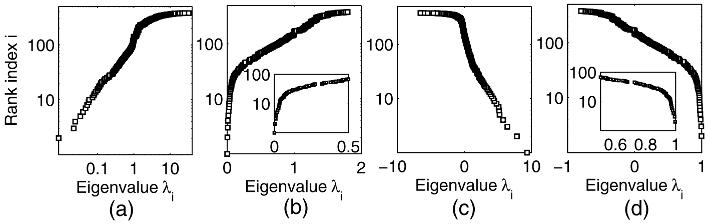


**Fig. 3.12** The spectrum of four matrices associated with the co-author network. The four matrices are respectively (**a**) the Laplacian matrix, (**b**) the normalized Laplacian matrix, (**c**) the modularity matrix, and (**d**) the normalized modularity matrix. The *horizontal axis* shows the eigenvalues of matrices and the *vertical axis* represents the rank index of eigenvalues. For the normalized Laplacian matrix and normalized modularity matrix, an *inset* is used to illustrate the largest eigengap. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

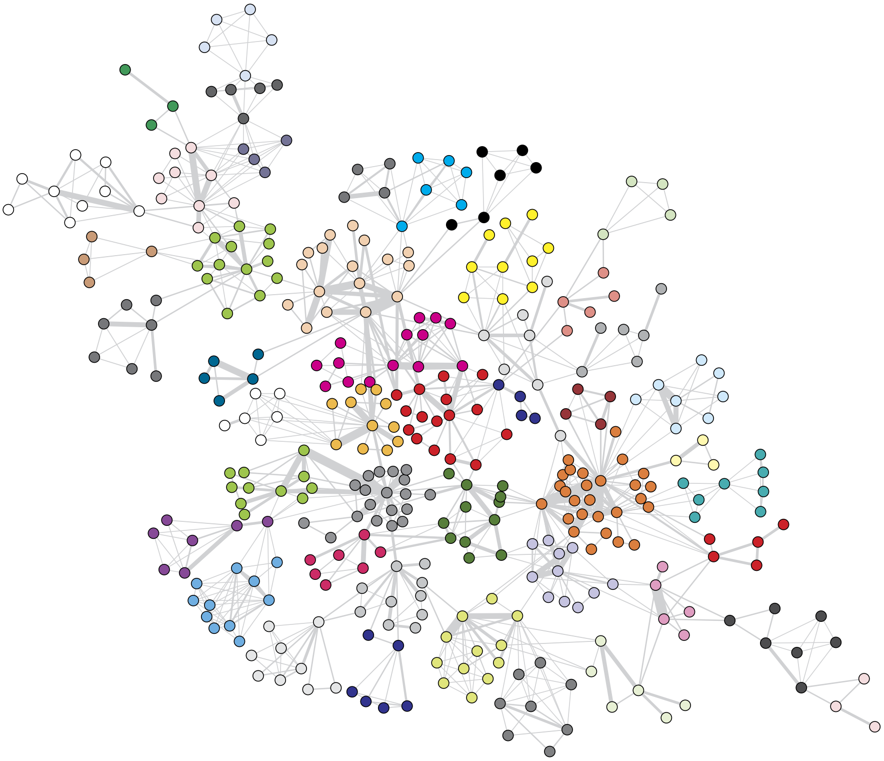


**Fig. 3.13** The topology of the co-author network. The width of each edge represents its weight. Different colors characterize the partition dividing the network nodes into 46 groups, which is obtained through using either of the two correlation matrices. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

### 3.5.2.2 Applications to More Real World Networks

Many tests on synthetic networks have demonstrated that the correlation matrices are superior to the covariance matrices at uncovering the intrinsic topological scale of networks. Now we apply the correlation matrices to more real world networks, which are widely used to evaluate community detection methods. These networks include the match network of the National Basketball Association teams in the 2009–2010 season, the journal index network constructed in [41], the social network of dolphins [22], the college football network of the United States [21], the metabolic network of *E. Coli* [42], the network of political books [31], the network of jazz musicians [43], the coauthor network of network scientists presented in [16], and the email network of University Rovira i Virgili [44]. For convenience, these networks are respectively abbreviated to *nba*, *journal*, *dolphin*, *football*, *ecoli*, *polbook*, *jazz*, *netsci* and *email*. The test results on these networks are shown in Fig. 3.14. Due to that the two correlation matrices achieve identical results, we only give the results of the normalized modularity matrix.

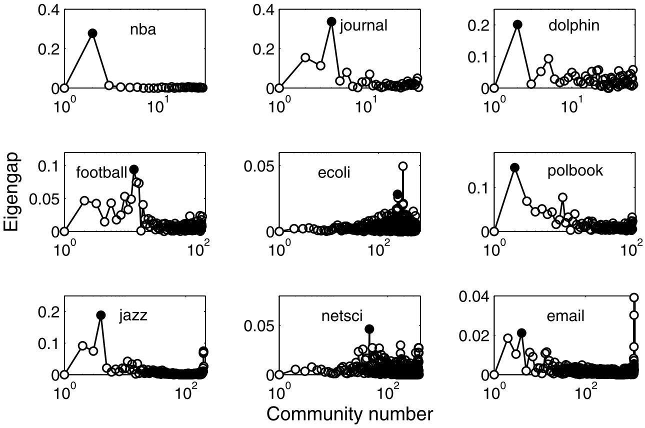


**Fig. 3.14** Applications to real world networks. The *vertical axis* represents the length of the eigengap and the *horizontal axis* represents the corresponding community number indicated by the eigengap. The *shaded-circles* mark the largest eigengap. Note that, for community detection, only the eigengaps between positive eigenvalues are taken into account. The *vertical dashed lines* are the place where the zero-valued eigenvalues occur. Reprinted from Ref. [33], with kind permission from Springer Science+Business Media

Figure 3.14 shows the occurrence of the largest eigengap of the correlation matrices. The corresponding communities reflect the structural and functional characteristics of each specific network, which can be verified by checking the nodes of each community. Specifically, for the networks with known community structure, including the networks *nba*, *journal*, *dolphin*, *football*, *ecoli* and *polbook*, our approach can accurately uncover such community structure. For the other three networks, the correlation matrices also give very promising results. In addition, for the last three networks, large eigengaps are observed among the negative eigenvalues of the normalized modularity matrix. This indicates that these networks contain anticommunity structure. Actually, this phenomenon is also observed in many other real world networks, which are not included in this thesis.

In addition, we also test the covariance matrices and the correlation matrices on some real world complex networks with larger size, including the protein interaction network and the word association network [45]. However, no significant eigengap is observed in the spectrum of the covariance matrix and the correlation matrix associated with these networks. This indicates that there is no scale which is significantly superior to other scales and thus no partition of network is more desired. However, unlike covariance matrices, the correlation matrices can characterize the cohesiveness of the communities at each specific scale through the magnitude of their eigenvalues. Generally speaking, an eigenvalue larger than 0.5 indicates the existence of a cohesive node group, i.e., a community. Thus, according to the magnitude of eigenvalues of the proposed correlation matrices, users can choose the topological scale and partition of networks based on their application needs.

## 3.6 Conclusions

In this section, we have studied the problem of detecting multiscale community structure in heterogeneous networks under a new framework of dimensionality reduction. This framework views community detection as a process of transforming a high-dimensional representation of a network, where each node is one dimension, to a low-dimensional representation, where each dimension corresponds to a community. This framework provides a unified explanation to the role of the Laplacian matrix for graph partitioning and the modularity matrix for community detection.

Based on the new framework, we first revealed that existing methods based on Laplacian and modularity matrices cannot effectively detect multiscale communities. We proposed to use the eigengaps of the covariance matrices to identify different topological scales. We then proposed a new method for detecting multiscale communities which uses significant eigenvectors to project network nodes into low-dimensional vectors and clusters those vectors into multiscale communities. Furthermore, we revealed that the Laplacian matrix and the modularity matrix cannot deal with the networks with heterogeneous node degrees. This problem is attributed to the fact that these two matrices only take into account the translation and rotation transformation. We proposed to use a rescaling transformation to handle heterogeneity. The correlation matrices resulted from the rescaling transformation are shown to be effective in detecting community structure for highly heterogeneous networks. Finally, we showed that although the Laplacian matrix and the modularity matrix behave very differently, the performance of their corresponding correlation matrices is almost identical. This further indicates that the rescaling transformation plays a critical role at the detection of multiscale community structure in heterogeneous networks.

Note that our framework is related to the eigenvalue decomposition and thus its scalability depends of development of the eigenvalue decomposition technique. However, real networks are usually very sparse and thus can alleviate the problem of scalability.

## References

1. Danon, L., Duch, J., Diaz-Guilera, A., Arenas, A.: Comparing community structure identification. J. Stat. Mech. P09008 (2005)
2. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
3. Leskovec, J., Lang, K.J., Mahoney, M.W.: Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International Conference on World Wide Web (WWW'10), pp. 631–640 (2010)
4. Lambiotte, R., Delvenne, J.C., Barahona, M.: Laplacian dynamics and multiscale modular structure in networks. arXiv:0812.1770 (2008)
5. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proc. Natl. Acad. Sci. USA **104**, 36–41 (2007)
6. Arenas, A., Fernández, A., Gómez, S.: Analysis of the structure of complex networks at different resolution levels. New J. Phys. **10**, 053039 (2008)

7. Ronhovde, P., Nussinov, Z.: Multiresolution community detection for megascale networks by information-based replica correlations. Phys. Rev. E **80**, 016109 (2009)
8. Arenas, A., Díaz-Guilera, A., Pérez-Vicente, C.J.: Synchronization reveals topological scales in complex networks. Phys. Rev. Lett. **96**, 114102 (2006)
9. Cheng, X.Q., Shen, H.W.: Uncovering the community structure associated with the diffusion dynamics on networks. J. Stat. Mech. P04024 (2010)
10. Delvenne, J.C., Yaliraki, S.N., Barahona, M.: Stability of graph communities across time scales. Proc. Natl. Acad. Sci. USA **107**, 12755–12760 (2010)
11. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community structure in time-dependent, multiscale, and multiplex networks. Science **328**, 876–878 (2010)
12. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**, 761–764 (2010)
13. Arenas, A., Borge-Holthoefer, J., Gómez, S., Zamora-López, G.: Optimal map of the modular structure of complex networks. New J. Phys. **12**, 053009 (2010)
14. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer Series in Statistics. Springer, New York (2002)
15. von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**, 395–416 (2007)
16. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**, 036104 (2006)
17. Fiedler, M.: Property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czechoslov. Math. J. **25**, 619–633 (1975)
18. Chung, F.R.K.: Spectral Graph Theory. Am. Math. Soc., Providence (1997)
19. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 888–905 (2000)
20. Ding, C.H.Q., He, X., Zha, H., Gu, M., Simon, H.D.: A min-max cut algorithm for graph partitioning and data clustering. In: Proceedings of the IEEE International Conference on Data Mining, pp. 107–114 (2001)
21. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**, 7821–7826 (2002)
22. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
23. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA **101**, 2658–2663 (2004)
24. Scott, J.: Social Network Analysis: A Handbook, 2nd edn. Sage, Thousand Oaks (2000)
25. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Trans. Knowl. Data Eng. **30**, 172–188 (2008)
26. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E **69**, 066133 (2004)
27. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. P10008 (2008)
28. Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Nature **433**, 895–900 (2005)
29. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Phys. Rev. E **72**, 027104 (2005)
30. Agarwal, G., Kempe, D.: Modularity-maximizing graph communities via mathematical programming. Eur. Phys. J. B **66**, 409–418 (2008)
31. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA **103**, 8577–8582 (2006)
32. Shen, H.W., Cheng, X.Q., Fang, B.X.: Covariance, correlation matrix, and the multiscale community structure of networks. Phys. Rev. E **82**, 016114 (2010)
33. Shen, H.W., Cheng, X.Q., Wang, Y.Z., Chen, Y.: A dimensionality reduction framework for detection of multiscale structure in heterogeneous networks. J. Comput. Sci. Tech. **27**, 341–357 (2012)
34. Newman, M.E.J.: Assortative mixing in networks. Phys. Rev. Lett. **89**, 208701 (2002)

35. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**, 452–473 (1977)
36. Chauhan, S., Girvan, M., Ott, E.: Spectral properties of networks with community structure. Phys. Rev. E **80**, 056114 (2009)
37. Capocci, C., Servedio, V.D.P., Caldarelli, G., Colaiori, F.: Detecting communities in large networks. Physica A **352**, 669–676 (2005)
38. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**, 046110 (2008)
39. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
40. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Phys. Rep. **424**, 175–308 (2006)
41. Rosvall, M., Bergstrom, C.T.: An information-theoretic framework for resolving community structure in complex networks. Proc. Natl. Acad. Sci. USA **104**, 7327–7331 (2007)
42. Muff, S., Rao, F., Caflisch, A.: Local modularity measure for network clusterizations. Phys. Rev. E **72**, 056107 (2005)
43. Gleiser, P., Danon, L.: Community structure in jazz. Adv. Complex Syst. **6**, 565–573 (2003)
44. Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F., Arenas, A.: Self-similar community structure in a network of human interactions. Phys. Rev. E **68**, 065103 (2003)
45. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)

# Chapter 4
# Community Structure and Diffusion Dynamics on Networks

## 4.1 Introduction

Community structure and network dynamics are two main research focuses in the study of complex networks. In the last decade, many methods for community detection have been proposed and applied successfully to some specific complex networks [1–7]. Meanwhile, network dynamics has also attracted much research attention [8, 9].

For community detection, each method requires, explicitly or implicitly, a definition of community from different perspectives, such as centrality measure, link density, percolation theory, and network compression. A well-known definition for community is modularity, which is proposed by Newman et al. as a quality function for a partition of network. Modularity is effective for detecting community structure of many real world networks. However, as pointed out by Fortunato et al. [10], modularity suffers the resolution limit problem and this problem raises concerns about the reliability of the communities detected through the optimization of modularity. In [11], the authors claimed that the resolution limit problem is attributed to the coexistence of multiple scale descriptions of the topological structure of network, while only one scale is obtained through directly optimizing the modularity. In addition, the definition of modularity only considers the significance of link density from the static topological structure of network, and it is unclear how the modularity based community structure is correlated to the dynamics on network.

For network dynamics, in recent years, researchers have begun to investigate the correlation between the community structure and the dynamics on networks. For example, Arenas et al. pointed out that the synchronization reveals the topological scale in complex networks [12]. In addition, the random walk on a network was also extensively studied and used to uncover community structure of the network [5, 13]. In [14, 15], the random walk on a network is introduced for defining the distance between network nodes, and an algorithm based on this distance is proposed for partitioning the network into communities. In [16], the authors proposed quantifying and ranking the quality of network partitions in terms of their stability, defined as the clustered autocovariance in the random walk process taking place on network.

H.-W. Shen, *Community Structure of Complex Networks*, Springer Theses,
DOI 10.1007/978-3-642-31821-4_4, © Springer-Verlag Berlin Heidelberg 2013

In this chapter, we study the relation between community structure and dynamics on networks by investigating the diffusion process taking place on network. We note that some local equilibrium states appear before the diffusion process reaches the final equilibrium state. The stability of these local equilibrium states can be measured by their duration time in the diffusion process. Then, we demonstrate that the intrinsic community structure is revealed by the stable local equilibrium states of the diffusion process. Furthermore, we show that such community structure can be directly identified through the optimization of network conductance, which measures how easily the diffusion among different communities occurs.

In addition, we show that the diffusion dynamics on network is closely correlated with the spectrum of the normalized Laplacian matrix. This inspires us to compare spectral methods with five different matrices in terms of their effectiveness at identifying the community structure of networks. The results of comparison demonstrate that the normalized Laplacian matrix and the normalized modularity matrix significantly outperform the other three unnormalized matrices at identifying the community structure of networks. This indicates that the heterogeneity of node degree is a crucial ingredient for the detection of community structure using spectral methods and the matrices that do not properly account for it are doomed to fail or to produce inaccurate results. Particularly, the modularity matrix does not gain desired benefits from using the configuration model as reference network with the consideration of the node degree heterogeneity.

## 4.2 Diffusion Dynamics on Networks

In this section, we describe the diffusion dynamics on networks. We first introduce some notations used later. An undirected network $G = (V, E)$ with $N$ nodes is often described in terms of its adjacency matrix $A$ whose elements $A_{xy}$ denote the strength of the link connecting nodes $x$ and $y$. The strength of node $x$ is denoted by $s_x = \sum_y A_{xy}$. For a node set $V_1 \subseteq V$, $|V_1|$ denotes the number of node in $V_1$, the volume of $V_1$ is defined as $vol(V_1) = \sum_{x \in V_1} s_x$, and $in\_vol(V_1) = \sum_{x \in V_1, y \in V_1} A_{xy}$ is referred to as the inward volume of $V_1$.

### 4.2.1 *Diffusion Process on Networks*

We start with investigating the diffusion process which describes the dynamics of a random walker moving on network. At each time $t$, the random walker moves from its current node $x$ to one of its neighboring nodes $y$ randomly with the probability $p(x \to y) = A_{xy}/s_x$. The dynamics of this process can be described as

$$\frac{d\rho_x(t)}{dt} = -r \sum_y L^T_{xy} \rho_y(t), \quad x = 1, \ldots, N, \tag{4.1}$$

where $\rho_x(t)$ is the probability that the random walker resides at the node $x$ at time $t$, and $r$ is a parameter controlling the rate of diffusion process. The matrix $L$ is the normalized Laplacian matrix defined as $L = I - D^{-1}A$, where $I$ is the identity matrix and $D$ is a diagonal matrix with its diagonal elements $D_{xx} = s_x$.

For any starting node, as time proceeds, the diffusion process described in Eq. 4.1 will definitely move towards equilibrium if the underlying undirected network is connected and non-bipartite [17]. When the diffusion process is at equilibrium state, it satisfies the so-called *detailed balance condition* [18], i.e., the probability that a random walker walks through nodes $x$ and $y$ successively is equal to the probability that this random walker walks through nodes $y$ and $x$ successively. Formally, the detailed balance condition can be denoted by $\rho_x(t)p(x \rightarrow y) = \rho_y(t)p(y \rightarrow x)$ and the reduced form is $\rho_x(t)/s_x = \rho_y(t)/s_y$ for undirected networks.

We explore all transients in the whole diffusion process instead of only the final equilibrium state. During the diffusion process on a network, it is known that the detailed balance condition is satisfied among highly interconnected nodes first and then, sequentially, among less interconnected ones, until among all the nodes. In order to evaluate how closely two nodes $x$ and $y$ satisfy the detailed balance condition at time $t$, we introduce a measure $c_{xy}(t)$ as

$$c_{xy}(t) = \left\langle \left| \frac{\rho_x(t)}{s_x} - \frac{\rho_y(t)}{s_y} \right| \right\rangle, \tag{4.2}$$

where $\langle \cdots \rangle$ averages over different realizations of the diffusion process with randomly selected starting nodes. In practice, a pair of nodes $x$ and $y$ is said to satisfy the detailed balance condition at the time $t$ when $c_{xy}(t)$ is smaller than a given threshold. A set $V$ of nodes is said to satisfy the detailed balance condition if the average value $\sum_{y \in V} c_{xy}(t)/|V|$ of $c_{xy}(t)$ for each node $x$ is smaller than the given threshold. Relative to the final equilibrium state, we say that the diffusion process is at a local equilibrium state when several groups of nodes locally satisfy the detailed balance condition. For convenience, we call the matrix $c_{xy}(t)$ as diffusion matrix. Using the diffusion matrix, we can trace the different local equilibrium states during the diffusion process.

As an example, we use $c_{xy}(t)$ to analyze the diffusion process on the H13-4 network, which is constructed according to Ref. [12]. This network has two predefined hierarchical levels. The first hierarchical level consists of 4 groups of 64 nodes and the second hierarchical level consists of 16 groups of 16 nodes. Figure 4.1 illustrates the diffusion matrix $c_{xy}(t)$ of two transients corresponding to two different local equilibrium states of the diffusion process. The squares along the diagonal suggest that the corresponding groups of nodes satisfy the detailed balance condition. These node groups reveal the predefined hierarchical levels in the H13-4 network. For comparison, we further investigate the diffusion dynamics on the randomized H13-4 network, which is constructed through shuffling the edges of the H13-4 network depicted in Fig. 4.1. From Fig. 4.2a, we can see that there is no node group locally satisfying the detailed balance condition.

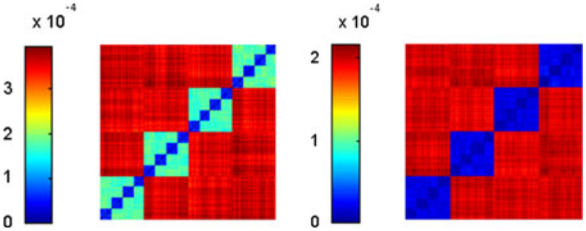


**Fig. 4.1** Diffusion matrix for two transients in the diffusion process on the H13-4 network. Here, each value of $c_{xy}(t)$ is the average over 10,000 realizations of the diffusion process with randomly selected starting nodes. The parameter $r = 0.01$. Reprinted from Ref. [19], Copyright 2010, with permission from IOP Publishing and SISSA

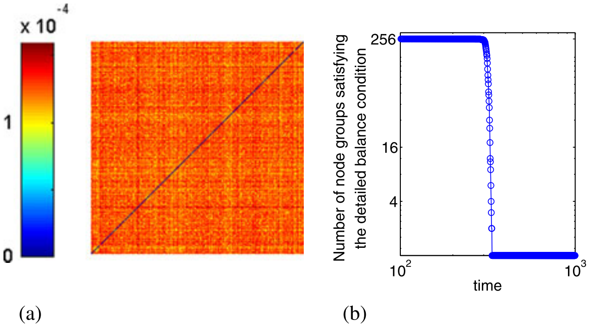


**Fig. 4.2** Diffusion dynamics on the randomized H13-4 network. (**a**) The matrix $c_{xy}(t)$ of a transient in the diffusion process. Each value of $c_{xy}(t)$ is the average over 10,000 realizations of the diffusion process with randomly selected starting nodes. The parameter $r = 0.01$. (**b**) The number of node groups satisfying the detailed balance condition as a function of time $t$. Here, the threshold for $c_{xy}(t)$ is set to be $1.0 \times 10^{-4}$. Reprinted from Ref. [19], Copyright 2010, with permission from IOP Publishing and SISSA

A phenomenon similar to what illustrated in Fig. 4.1 has also been observed in the synchronization process. In [12], the authors claimed that this phenomenon reveals the topological scale of networks. The authors also pointed out that local equilibrium state phenomenon in synchronization process is correlated with the spectrum of the Laplacian matrix associated with the underlying network. According to the characteristics of Laplacian matrix, as pointed out in [20], the community structure revealed by synchronization process is heavily affected by the heterogeneous distributions of degree and community size. In the following, we will show that the local equilibrium state phenomenon is correlated with the spectrum of normalized Laplacian matrix, which takes the heterogeneous degree and community size distribution into account. Thus, the normalized Laplacian matrix outperforms the Laplacian matrix at clustering the nodes of network [20].

A local equilibrium state is regarded as stable if the set of node groups satisfying the detailed balance condition remains unchanged for a long duration in diffusion process. To investigate the stability of local equilibrium states, we study the solution of Eq. 4.1 in terms of the normal modes $\varphi_i(t)$, which reads

$$\rho_x(t) = \sum_i U_{xi}\varphi_i(t) = \sum_i U_{xi}\varphi_i(0)e^{-\lambda_i rt}, \quad x = 1, \ldots, N, \tag{4.3}$$

where $\lambda_i$ are the eigenvalues of the transpose of normalized Laplacian matrix $L$, and $U$ is the eigenvector matrix whose $i$th column is the eigenvector $u_i$ corresponding to eigenvalue $\lambda_i$. Given the starting node of diffusion process, the initial amplitudes $\varphi_i(0)$ can be determined according to Eq. 4.3 due to the eigenvector matrix $U$ being fixed, i.e., $\varphi_i(0)$ only depends on the starting node of diffusion process. Note that, as pointed out in Eq. 4.2, we investigate the average behavior of many different diffusion processes with randomly selected starting nodes. Thus, the choice of starting nodes does not affect the analysis results. Without loss of generality, we rank these eigenvalues in the ascending order $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_i \leq \cdots \leq \lambda_N$.

As time proceeds in diffusion process, these normal modes $\varphi_i(t) = \varphi_i(0)e^{-\lambda_i rt}$ $(i \neq 1)$ will decay to zero. We use $\tau_i$ to denote the time when the normal mode $\varphi_i(t)$ decays to zero. Formally, $\tau_i$ is infinite. In practice, a threshold $\varepsilon$ is usually used to determine when $\varphi_i(t)$ decays to zero, i.e., $\varphi_i(t) < \varepsilon$. In this case, we have

$$\tau_i = \frac{1}{\lambda_i} \times \frac{\ln \varphi_i(0) - \ln \varepsilon}{r}. \tag{4.4}$$

All these moments $\tau_i$ $(1 \leq i \leq N)$ form a series of time intervals, respectively $[\tau_{N+1} = 0, \tau_N), [\tau_N, \tau_{N-1}), \ldots, [\tau_{i+1}, \tau_i), \ldots, [\tau_3, \tau_2), [\tau_2, \tau_1 = \infty)$. These time intervals divide the whole diffusion process into $N$ stages. Specifically, the time interval $[\tau_{i+1}, \tau_i)$ is regarded as the $i$th stage. When the diffusion process is at the $i$th stage, only the normal modes $\varphi_j(t)$ $(1 \leq j \leq i)$ have not decayed to zero. Thus we have

$$\rho_x(t) \approx \sum_{j=1}^{i} U_{xj}\varphi_j(t), \quad x = 1, \ldots, N. \tag{4.5}$$

This indicates that the value $\rho_x(t)$ of node $x$ at the $i$th stage can be represented by the $i$-dimension coefficient vector of $\varphi_j(t)$, i.e., $(U_{x1}, U_{x2}, \ldots, U_{xj}, \ldots, U_{xi})$. According to Eqs. 4.2 and 4.5, given a threshold, we can identify the node groups satisfying detailed balance condition through clustering the normalized $i$-dimension vectors of $(U_{x1}, U_{x2}, \ldots, U_{xj}, \ldots, U_{xi})$ using, for example, the $k$-means clustering method. The set of such node groups is unchanged due to the non-decayed normal modes being fixed during the same stage. This means that a local equilibrium state is stable if the corresponding stage persists for a long time.

Taking the H13-4 network as an example, the left panel of Fig. 4.3 shows the different local equilibrium states of diffusion process and the right panel illustrates the different stages in terms of the number of non-decayed normal modes of diffusion process. Through comparing the two panels, we see that each time one normal mode decays to zero, the diffusion process changes from a local equilibrium state to a new one. It is observed that two stable local equilibrium states with long durations emerge in diffusion process. The node groups corresponding to these two

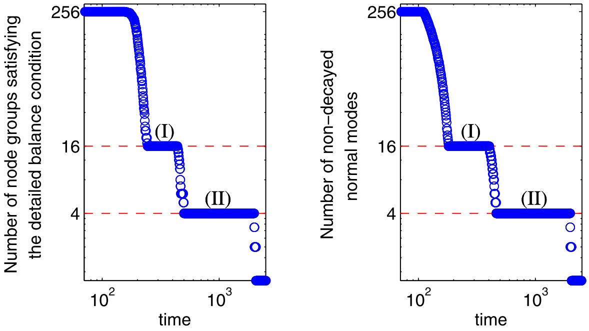


**Fig. 4.3** Relation between node groups and non-decayed normal modes. *Left panel*: the number of node groups satisfying the detailed balance condition as a function of time $t$. Here, the threshold for $c_{xy}(t)$ is set to be $1.0 \times 10^{-4}$. *Right panel*: the number of non-decayed normal modes in terms of the time $t$. Reprinted from Ref. [19], Copyright 2010, with permission from IOP Publishing and SISSA

states clearly reveal the intrinsic community structure of H13-4 network. In addition, from Fig. 4.2b, we can see that no stable local equilibrium state appears in the diffusion process taking place on the randomized H13-4 network. This is reasonable since it is commonly believed that randomized network have no community structure. All these findings suggest that the appearance of stable local equilibrium states in a diffusion process indicates the existence of community structure in the underlying network.

### *4.2.2 Network Conductance and Community Structure*

Note that diffusion occurs much more frequently within node groups than among them when the diffusion process is at a stable local equilibrium state. This indicates that there exists a high transitive cohesion inside such node groups. The community structure comprised of these node groups could well reflect the diffusion dynamics on the underlying network. Regarding such community structure as a partition of a network, we measure the quality of the partition through introducing the *conductance* of network, which reflects how easily the diffusion occurs among different communities.

For a given partition $\mathscr{P} = \{V_1, \ldots, V_k\}$, conductance is defined as the average *departure probability*, $p_{dept}(V_i)$, of all the communities $V_i$, that is $C(\mathscr{P}) = \frac{1}{k}\sum_{i=1}^{k} p_{dept}(V_i)$. The departure probability of a community $V_i$ is the probability that a random walker departs from $V_i$ in the next time step given that it resides at $V_i$ when the diffusion process is at the final equilibrium state. Formally, the departure probability of $V_i$ can be computed by using

$$p_{dept}(V_i) = \frac{\sum_{x \in V_i, y \in \overline{V}_i} \rho_x(\infty) p(x \rightarrow y)}{\sum_{x \in V_i} \rho_x(\infty)} = 1 - \frac{in\_vol(V_i)}{vol(V_i)}, \qquad (4.6)$$

where $\rho_x(\infty) = s_x/vol(V)$ is the stationary distribution which characterizes the final equilibrium state of the diffusion process. In this way, the conductance is formally denoted by

$$C(\mathscr{P}) = \frac{1}{k}\sum_{i=1}^{k}\left(1 - \frac{in\_vol(V_i)}{vol(V_i)}\right). \tag{4.7}$$

Actually, it can be proved that the community structure associated with the stable local equilibrium state can be exactly identified through minimizing the conductance directly. Without loss of generality, we assume that the stable local equilibrium state emerges at the $k$th stage in the diffusion process. As mentioned above, the community structure associated with this state can be identified through clustering the normalized $k$-dimension vectors of $(U_{x1}, U_{x2}, \ldots, U_{xi}, \ldots, U_{xk})$. Further, through the matrix trace maximization method [21], it can be proved that the optimization of the conductance for a fixed $k$ can be done through clustering the top $k$ eigenvectors of the transpose of the normalized Laplacian matrix, corresponding to the 1st to $k$th columns of $U$. Therefore, the optimization of conductance provides an effective way to identify the community structure associated with the stable local equilibrium state.

Now we clarify the difference between the conductance and the earlier measure for the quality of network partition from the perspective of a random walk on networks. Firstly, as pointed out in [22], as a measure of the quality of network partition, the modularity can be described as the difference between the probability that a random walker resides in the same community on two successive time steps and the probability that two independent random walkers both resides in the same community. Secondly, in [16], through considering the random path with length $t$ instead of the length 1 for the modularity and the length of infinity for the spectral partition, the authors proposed that the stability of network partitions be defined as the clustered autocovariance of the random walk. This stability provides a general framework for quantifying and ranking the quality of network partitions from the perspective of the whole network, i.e., it characterizes the fraction of the within-community random paths with length $t$ with respect to all the random paths of length $t$. However, the conductance considers the quality of network partition from the perspective of each community instead of the whole network, i.e., it reflects the fraction of within-community random paths with respect to all the random paths departing solely from the community considered. This provides the advantage for handling the heterogeneous distribution of community size (or volume) which is common to real world networks. As follows, the application of the conductance optimization method to the benchmarks of Lancichinetti et al. also demonstrates that our method can effectively handle the heterogeneous distribution of community size.

To test the effectiveness of our method for community detection based on the optimization of conductance, we utilize the benchmark proposed by Lancichinetti et al. in [23]. This benchmark provides networks with heterogeneous distributions of node degree and community size. Thus it poses a much more severe test of community detection algorithms than standard benchmarks. Many parameters are used

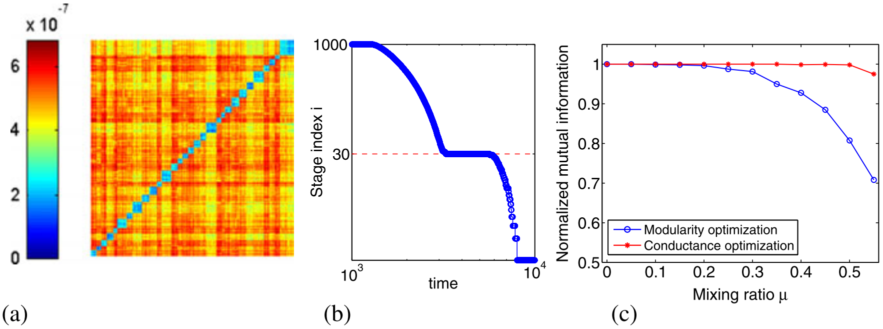


**Fig. 4.4** Experimental results on a benchmark network. For the benchmark network, the mixing ratio $\mu = 0.3$ and the number of communities is 30. (**a**), (**b**) The most stable local equilibrium and the different stages of the diffusion process. (**c**) Comparison between the conductance optimization method and the modularity optimization method on benchmark networks. Each point corresponds to an average over 100 network realizations. Reprinted from Ref. [19], Copyright 2010, with permission from IOP Publishing and SISSA

to control the generated networks in this benchmark: the number of nodes $N$, the average node degree $\langle k \rangle$, the maximum node degree max_$k$, the mixing ratio $\mu$, the exponent of the power law node degree distribution $t_1$, the exponent of the power law distribution of community size $t_2$, the minimum community size min_$c$, and the maximum community size max_$c$. In our tests, we use the default parameter configuration where $N = 1000$, $\langle k \rangle = 15$, max_$k = 50$, $t_1 = 2$, $t_2 = 1$, min_$c = 20$, and max_$c = 50$. By tuning the parameter $\mu$, we test the effectiveness of our method on the networks with different fuzziness of communities. The larger the parameter $\mu$, the fuzzier the community structure of the generated network. In addition, we adopt the normalized mutual information (NMI) [24] in order to compare the partition found by the algorithms with the answer partition. The larger the NMI is, the more effective the tested algorithm.

Figure 4.4a–b illustrate the most stable local equilibrium state and the different stages of the diffusion process on the benchmark network with the mixing ratio $\mu = 0.3$ and the number of communities equal to 30. The squares along the diagonal indicate the predefined communities in the network. The number of these communities is clearly revealed by the most stable local equilibrium state. Figure 4.4c shows the comparison between the conductance optimization method and the modularity optimization method in terms of the NMI on the benchmark network. When the community structure is evident, both our method and the modularity optimization method (e.g., the fast unfolding algorithm [6] and the spectral method [4]) can accurately identify the community structure. However, when the community structure becomes fuzzier, the performance of the modularity optimization method deteriorates while our method still achieves rather good results.

In addition, we also tested the conductance optimization method on many real world networks which are widely used to evaluate community detection methods. These networks include the social network of Zachary's karate club [25], the social network of dolphins of Lusseau et al. [26], the college football network of the United

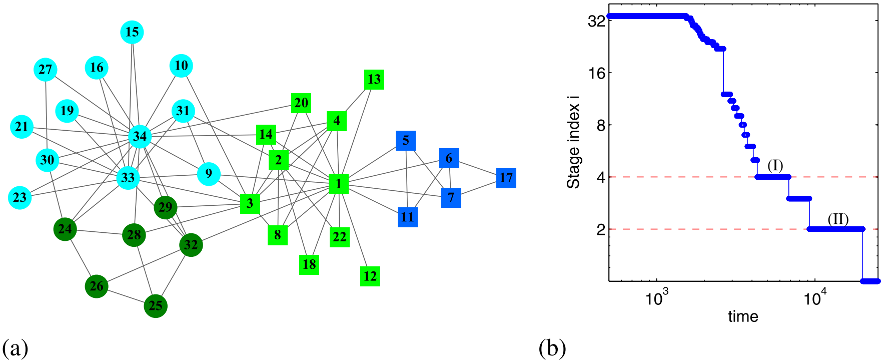


**Fig. 4.5** Illustration of the conductance optimization method on a real world network. (**a**) The friendship network of the karate club. Colors are used to differentiate the communities uncovered by the conductance optimization method when considering the stage (I). Shapes, circle and square, are used to distinguish the communities corresponding to the stage (II) and the real social split of this network observed by Zachary. (**b**) The different stages of the diffusion process taking place on the karate club network. Two most stable equilibrium states are marked with (I) and (II). Reprinted from Ref. [19], Copyright 2010, with permission from IOP Publishing and SISSA

States [1], the journal index network constructed in [5], and the network of political books [4]. For all these networks, the conductance optimization method obtains extremely good results. Taking Zachary's network as an example, Fig. 4.5b illustrates the different stages of the diffusion process taking place on the network. The two most stable local equilibrium states are marked (I) and (II), and the corresponding communities are depicted in Fig. 4.5a. Besides the stages (I) and (II), another relatively stable state can be also observed during the diffusion process, as shown in Fig. 4.5b. The corresponding three communities are respectively the one comprised of all the circle nodes and two communities formed by the square nodes but with different colors, as shown in Fig. 4.5a. Actually, as regards the three stable states, it is really hard to say which the best one is. The duration time of each state may provide an effective candidate measure for the significance of network divisions.

## 4.3 Comparative Analysis of Spectral Methods for Community Detection

In the previous section, we have shown that the spectrum of normalized Laplacian matrix provides critical indicator for the detection of community structure associated with the diffusion dynamics. This inspires us to study the general spectral method for community detection. In this section, we will give a comparative analysis of the spectral methods based on five different matrices, namely adjacency matrix, standard Laplacian matrix, normalized Laplacian matrix, modularity matrix and correlation matrix.

### 4.3.1 The Matrices for Spectral Analysis

The topology of network is often described in terms of adjacency matrix. Based on the adjacency matrix, several other matrices are formulated to investigate the properties of network, including the standard Laplacian matrix, the normalized Laplacian matrix, the modularity matrix and the correlation matrix. Existing studies indicate that the spectrum of these matrices sheds light on the community structure of network. In the following, we first give the definition of these matrices and briefly introduce the methods to detect the community structure using the spectrum of these matrices.

- *Adjacency matrix.* The elements $A_{ij}$ of an adjacency matrix $A$ denote the strength of the edge connecting the nodes $i$ and $j$ if such an edge exists, and 0 otherwise. (We restrict our attention in this paper to undirected networks.) In [27], the authors proposed that the spectrum of the adjacency matrix can unravel the number of communities. Specifically, the eigenvalues of the adjacency matrix is ranked in descending order, i.e., $\lambda_1^A \geq \lambda_2^A \geq \cdots \geq \lambda_i^A \geq \cdots \geq \lambda_n^A$, where $n$ is the number of network nodes. Each two successive eigenvalues form an eigengap, the $i$th eigengap being between $\lambda_i^A$ and $\lambda_{i+1}^A$ $(1 \leq i \leq n-1)$. The length of the $i$th eigengap is $\lambda_i^A - \lambda_{i+1}^A$. Then, the number of communities is indicated by the place of the largest eigengap, i.e., $i$ is the number of communities if the largest eigengap is the $i$th one.
- *Standard Laplacian matrix.* The standard Laplacian matrix is defined as $L = D - A$, where $D$ is a diagonal matrix with the diagonal element $D_{ii}$ being the degree of the node $i$. As to the standard Laplacian matrix, the Fiedler's vector [28] has been well studied and widely used for two-way network partition. Fiedler's vector is the eigenvector of the standard Laplacian matrix corresponding to the second smallest eigenvalue. More importantly, the standard Laplacian matrix is often used to characterize the synchronization dynamics on networks [12, 29]. In [12], Arenas et al. pointed out that the spectrum of the standard Laplacian matrix reveals the intrinsic topological scales. The eigenvalues are ranked in ascending order and the length of the $i$th eigengap is defined as $\log \lambda_{i+1}^L - \log \lambda_i^L$ $(2 \leq i \leq n-1)$.[1] Then, $i$ is viewed as the appropriate candidate for the number of intrinsic communities if the $i$th eigengap is largest.
- *Normalized Laplacian matrix.* The normalized Laplacian matrix is often defined as $N = I - T$, where $I$ is the identity matrix and $T$ is the transition matrix, which is defined as $T = D^{-1}A$ with the elements $T_{ij}$ being the probability that a random walker moves to the node $j$ from the node $i$. The normalized Laplacian matrix is named after the fact that it can be written in the form $N = D^{-1}L$, i.e., normalizing the standard Laplacian matrix with the diagonal matrix $D$ of node degrees. In [30], the authors claimed that the spectrum of the transition matrix $T$ can be

[1] In this paper, we also tested the alternative eigengap defined as $\lambda_{i+1}^L - \lambda_i^L$ $(1 \leq i \leq n-1)$, and the results are similar.

used to detect the community structure of networks. Actually, if $\lambda$ is an eigenvalue of the transition matrix, $1 - \lambda$ is an eigenvalue of the normalized Laplacian matrix with the same eigenvector. Furthermore, the normalized Laplacian matrix is closely correlated to the diffusion dynamics on networks. Through investigating the diffusion dynamics on networks, Cheng and Shen pointed out [19] that the community structure can be identified through the eigenvalues and eigenvectors of the normalized Laplacian matrix. Specifically, the eigenvalues are ranked in ascending order and the length of the $i$th eigengap is defined as $\lambda_{i+1}^N - \lambda_i^N$ $(1 \le i \le n-1)$. Then, $i$ is viewed as the appropriate candidate for the number of intrinsic communities if the $i$th eigengap is largest.

- *Modularity matrix*. The modularity matrix is proposed by Newman as a spectral explanation for the well-known measure, namely modularity, for the quality of network partition [4, 31]. Its elements are defined as

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m},$$

where $k_i = \sum_j A_{ij}$ is the strength of the node $i$ and $2m = \sum_{ij} A_{ij} = \sum_i k_i$ is the total strength of all the nodes. In [31], the eigenvectors corresponding to positive eigenvalues are utilized to uncover the community structure of networks. The number of communities can be determined according to the magnitude of the positive eigenvalues. Here, we rank the eigenvalues in descending order and the length of the $i$th eigengap is defined as $\lambda_{i-1}^B - \lambda_i^B$ $(2 \le i \le n)$. Then, $i$ is taken as the number of communities if the $i$th eigengap has the largest length. Note that, for the purpose of the detection of community structure, only the eigengaps among positive eigenvalues are considered. If all the eigenvalues are negative, no natural community structure exists, i.e., all the nodes belong to a sole community and the community number is 1. In [32], the modularity matrix is shown to be the biased covariance matrix of network and the spectrum of the covariance matrix is investigated for the detection of the multiscale community structure.

- *Correlation matrix*. The correlation matrix of network characterizes the correlation coefficients between pairs of nodes. Its element $C_{ij}$ are defined as

$$C_{ij} = \frac{B_{ij}}{\sqrt{k_i - k_i^2/2m}\sqrt{k_j - k_j^2/2m}}.$$

In [32], the correlation matrix is used to uncover the multiscale community structure of networks. Specifically, the eigenvalues are ranked in descending order and the length of the $i$th eigengap is defined as $\lambda_{i-1}^C - \lambda_i^C$ $(2 \le i \le n)$. The same to the modularity matrix, only the eigengaps among positive eigenvalues are considered. Then, $i$ is taken as the number of communities if the $i$th eigengap has the largest length. A similar matrix is called the symmetric normalized Laplacian matrix, whose element at the place $(i, j)$ is defined as $\delta_{ij} - A_{ij}/\sqrt{k_i}\sqrt{k_j}$, where $\delta_{ij}$ is 1 when $i = j$ and 0 otherwise. This matrix is often used in spectral clustering algorithms together with the two aforementioned Laplacian matrices [20].

In summary, the number of communities can be determined according to the eigengaps of the aforementioned five matrices. Actually, the community structure can be further identified using the eigenvectors of these matrices. Generally speaking, only several eigenvectors are utilized to project each node into a low-dimensional node vectors, and then the community structure is identified through clustering the node vectors using, for example, the $k$-means clustering method. Specifically, the selected eigenvectors correspond to the largest $n_c$ eigenvalues for the adjacency matrix, the smallest $n_c$ eigenvalues for the standard Laplacian matrix and the normalized Laplacian matrix, the largest $n_c - 1$ eigenvalues for the modularity matrix and the correlation matrix. Here, $n_c$ is the number of communities. These selected eigenvectors are stacked as columns of a matrix and the transpose of the $i$th row of this matrix is taken as the projected node vector corresponding to the node $i$. The community structure is then detected through clustering the projected node vectors.

Before proceeding, we first clarify why we choose the general method for community detection using the $k$-means clustering method. On one hand, this paper only considers the performance of the aforementioned five matrices rather than the specific implementation of spectral methods. Thus, it is fair and reasonable to choose the general method, which determines the community number according to the spectrum of matrices and identifies the community structure using the eigenvectors of the matrices. On the other hand, the $k$-means clustering method is the common practice for the spectral clustering [20]. Furthermore, the $k$-means clustering method is facilitated by the projected node vector subspace spanned by the top eigenvectors of matrices [33]. Thus, using the $k$-means clustering on the node vectors provides a competitive candidate among all the spectral methods for the detection of community structure.

Now, as an example, we illustrate the spectral methods through application on the Zachary's karate club network, which has been widely used to evaluate the community detection methods. This network characterizes the social interactions between the individuals in a karate club at an American university. A dispute arose between the club's administrator and its principal karate teacher, and as a result the club eventually split into two smaller clubs, centered around the administrator and the teacher respectively. The network and its fission is depicted in Fig. 4.5a. The administrator and the teacher are represented by nodes 1 and 33 respectively. Figure 4.6 shows the spectrum of the aforementioned five matrices associated with the Zachary's karate club network. The largest eigengap of the adjacency matrix, the standard Laplacian matrix and the modularity matrix indicate that the optimal number of community is 2. The corresponding community structure is consistent with the real split of the network. However, as indicated by the largest eigengap of the normalized Laplacian matrix and the correlation matrix, 4 is the optimal number of communities. The corresponding four communities are shown in Fig. 4.5a differentiated with colors, which is the results of many existing methods for community detection including the modularity maximization.

As illustrated by the previous example, the five matrices give rise to two different resulting partitions as the community structure of the network. Actually these

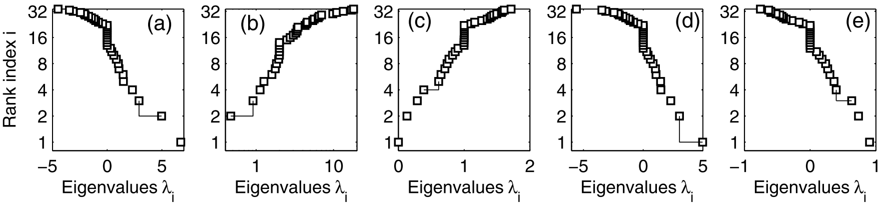


**Fig. 4.6** The spectrum of five matrices associated with the Zachary's karate club network. This five matrices are respectively (**a**) the adjacency matrix, (**b**) the standard Laplacian matrix, (**c**) the normalized Laplacian matrix, (**d**) the modularity matrix and (**e**) the correlation matrix. For each matrix, the largest eigengap is marked with an elbow line. Reprinted from Ref. [34], Copyright 2010, with permission from IOP Publishing and SISSA

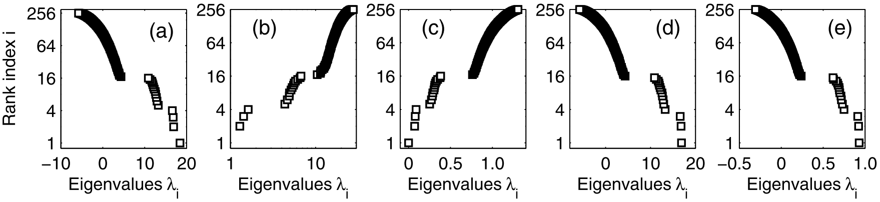


**Fig. 4.7** The spectrum of five matrices associated with the H13-4 network. The five matrices are respectively (**a**) the adjacency matrix, (**b**) the standard Laplacian matrix, (**c**) the normalized Laplacian matrix, (**d**) the modularity matrix and (**e**) the correlation matrix. Reprinted from Ref. [34], Copyright 2010, with permission from IOP Publishing and SISSA

two partitions correspond to two different topological scales of the network. The multiple scale of topological description is a common phenomenon in real-world networks [10, 11, 32, 35–37]. Actually, the multiscale community structure can be revealed through considering more eigengaps besides the largest one among the eigenvalues of the aforementioned five matrices. As an example, we illustrate the detection of the multiscale community structure of the H13-4 network, which is constructed according to [12]. This network has two predefined hierarchical levels. The first hierarchical level consists of 4 groups of 64 nodes and the second hierarchical level consists of 16 groups of 16 nodes. On average, each node has 13 edges connecting to the nodes in the same group at the second hierarchical level and has 4 edges connecting to the nodes in the same group at the first hierarchical level. This explains the name of such kind of networks. In addition, the average degree of each node is 18. According to the construction rules of the H13-4 network, the two hierarchical levels constitute the different topological descriptions of the community structure of the H13-4 network at different scales. As shown in Fig. 4.7, the community numbers associated with the two predefined topological scales are clearly revealed by the top two largest eigengaps occurring in the spectrum of the five matrices. The resulting communities are exactly the predefined node groups under the two hierarchical levels. However, according to the length of eigengap, the standard Laplacian matrix seems to prefer the first hierarchical level while the other four matrices tend to reveal the second hierarchical level.

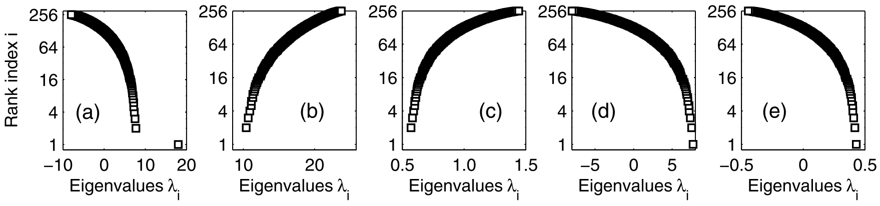


**Fig. 4.8** The spectrum of five matrices associated with the randomized H13-4 network. The five matrices are respectively (**a**) the adjacency matrix, (**b**) the standard Laplacian matrix, (**c**) the normalized Laplacian matrix, (**d**) the modularity matrix and (**e**) the correlation matrix. Reprinted from Ref. [34], Copyright 2010, with permission from IOP Publishing and SISSA

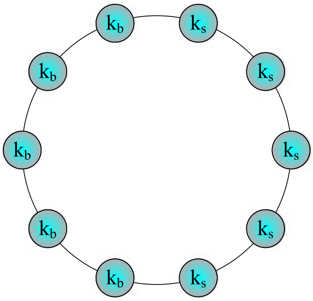


**Fig. 4.9** The clique circle network as a schematic example. Each *circle* corresponds to a clique, whose size is marked by its label $k_s$ or $k_b$. Reprinted from Ref. [34], Copyright 2010, with permission from IOP Publishing and SISSA

Furthermore, we apply all these matrices to the random network. For comparison, we construct the random network through shuffling the edges of the H13-4 network. Figure 4.8 shows the spectrum of the five matrices associated with the randomized H13-4 network. The spectrum of these matrices indicates that the number of communities is 1 or 256, i.e., all the nodes belong to the same community or each node forms a community. These findings are reasonable since it is commonly believed that randomized networks have no community structure.

The previous examples show that the aforementioned five matrices are both effective at revealing the community structure of network. Note that, as to the example H13-4 network, the nodes have approximately the same degree and the communities at a specific scale are of the same size. However, the real world networks usually have heterogeneous distributions of node degree and community size. Thus it will be more convincing to test these matrices on networks with heterogeneous distributions of node degree and community size. Before we give such a test in the subsequent section, using a schematic network, we first illustrate the difference between the effectiveness of these matrices. The schematic network is often called the clique circle network as depicted in Fig. 4.9. Generally speaking, the intrinsic community structure corresponds to the partition where each clique is taken as a community, which is the sole intrinsic scale existing in this network. As shown in Fig. 4.10, the sole topological scale is exactly revealed by the spectrum of the standard Laplacian matrix, the normalized Laplacian matrix and the correlation matrix. However, two scales

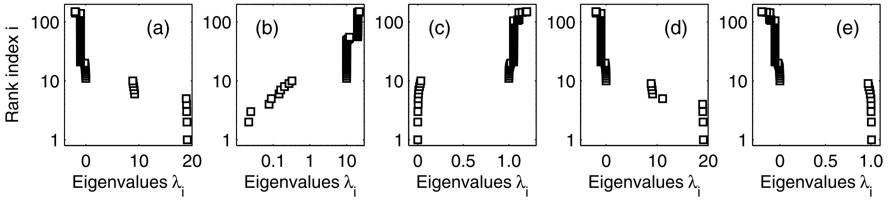


**Fig. 4.10** The spectrum of five matrices on the clique circle network with $k_s = 10$ and $k_b = 20$. These five matrices are respectively (**a**) the adjacency matrix, (**b**) the standard Laplacian matrix, (**c**) the normalized Laplacian matrix, (**d**) the modularity matrix and (**e**) the correlation matrix. Reprinted from Ref. [34], Copyright 2010, with permission from IOP Publishing and SISSA

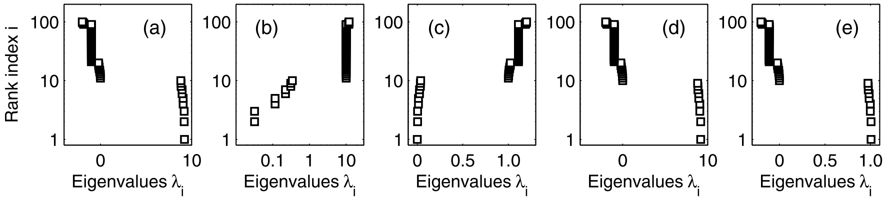


**Fig. 4.11** The spectrum of five matrices on the clique circle network with clique size $k_s = k_b = 10$. These matrices are respectively (**a**) the adjacency matrix, (**b**) the standard Laplacian matrix, (**c**) the normalized Laplacian matrix, (**d**) the modularity matrix and (**e**) the correlation matrix. Reprinted from Ref. [34], Copyright 2010, with permission from IOP Publishing and SISSA

are observed when we investigate the community structure of this network using the spectrum of the adjacency matrix and the modularity matrix. One scale corresponds to the intrinsic scale of the network, and the other corresponds to the partition dividing the network nodes into 5 groups, which is not desired. In [10], Fortunato et al. pointed out the resolution limit problem of the modularity through investigating the modularity maximization on such a clique circle network with each clique having the same size. However, as shown in Fig. 4.11, when all the cliques have the same size (i.e., the homogeneous node degree), the intrinsic community structure can be exactly revealed by all the five matrices, including the modularity matrix. This indicates that the resolution limit problem of the modularity is not the same to the problem studied in this paper. Specifically, the resolution limit problem means that there exists an intrinsic scale beyond which the smaller communities cannot be detected through maximizing the modularity. As to the heterogeneity problem of the modularity matrix considered in this paper, we focus on whether the modularity matrix can reveal the natural community structure, which can be detected using the spectral clustering method instead of the modularity maximization. In sum, the resolution limit problem talks about the maximization of modularity while the heterogeneity problem takes root in the modularity matrix. Thus we claim that it is crucial to deal with the heterogeneous degree when using the spectral methods for community detection.

### 4.3.2 Tests on Benchmark Networks

In this section, we show the effectiveness of the aforementioned five matrices at identifying the community structure on benchmark networks. We utilize the benchmark proposed by Lancichinetti et al. [23]. This benchmark provides networks with heterogeneous distributions of node degree and community size. Thus it poses a much more severe test to community detection algorithms than Newman's standard benchmark [2]. Many parameters are used to control the generated networks in this benchmark: the number of nodes $N$, the average node degree $\langle k \rangle$, the maximum node degree max_$k$, the mixing ratio $\mu$, the exponent $\gamma$ of the power law distribution of node degree, the exponent $\beta$ of the power law distribution of community size, the minimum community size min_$c$, and the maximum community size max_$c$. In our tests, we use the default parameter configuration where $N = 1000$, $\langle k \rangle = 15$, max_$k = 50$, min_$c = 20$, and max_$c = 50$. To test the influence from the distribution of node degree and community size, we adopt four parameter configurations for $\gamma$ and $\beta$, respectively being $(\gamma, \beta) = (2, 1)$, $(\gamma, \beta) = (2, 2)$, $(\gamma, \beta) = (3, 1)$ and $(\gamma, \beta) = (3, 2)$. Finally, by tuning the parameter $\mu$, we test the effectiveness of the five matrices on the networks with different fuzziness of community structure. The larger the mixing ratio parameter $\mu$, the fuzzier the community structure of the generated network.

The first test focuses on whether the number of communities can be correctly identified. Note that each benchmark network has only one significant topological scale according to the construction rules. Thus we only consider whether such a scale can be revealed by the largest eigengap in the spectrum of the five matrices. For each given mixing ratio $\mu$, 100 benchmark networks are generated. For each network, we use the spectrum of the aforementioned five matrices to identify the number of communities. The performance of each method is characterized by the fraction of benchmark networks whose community number is correctly identified. As shown in Fig. 4.12, the best results are obtained by the methods based on the normalized Laplacian matrix and the correlation matrix, which actually give the identical results for all the four used parameter configurations. When the mixing ratio $\mu$ is smaller than 0.5, i.e., the communities are defined in the strong sense [38], the number of communities can be accurately identified by investigating the spectrum of the normalized Laplacian matrix or the correlation matrix. Even when $\mu$ is larger than 0.5 (e.g., 0.55), these two matrices still give rather good results. The adjacency matrix and the modularity matrix exhibit rather similar effectiveness, obtaining very good results when the community structure is evident and deteriorating when the community becomes fuzzier with the increase of the mixing ratio $\mu$. Compared with the other four matrices, the standard Laplacian matrix gives the worst results, failing to identify the correct number of communities even when the community structure is quite evident. In addition, the exponent $\gamma$ of the power law distribution of node degree affects the effectiveness of the matrices except the normalized Laplacian matrix and the correlation matrix. The possible reason is that these two matrices take into account the distribution of node degree through the normalization operation in their definition. Finally, as shown in Fig. 4.12, it seems that all these five matrices

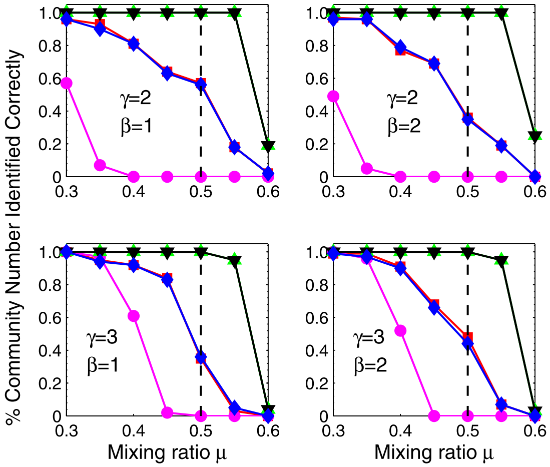


**Fig. 4.12** Effectiveness comparison at identifying community number. The benchmark networks are generated with four different parameter configurations. For each parameter configuration, 100 generated networks are used. The corresponding matrices are respectively the adjacency matrix (□), the standard Laplacian matrix (◯), the normalized Laplacian matrix (△), the modularity matrix (◊) and the correlation matrix (▽). Reprinted from Ref. [34], Copyright 2010, with permission from IOP Publishing and SISSA

are not very sensitive to the exponent $\beta$ of the power law distribution of community size.

The second test turns to the performance of the eigenvectors of the five tested matrices. Given the number of communities, we investigate whether the predefined community structure can be identified using the eigenvectors of the five tested matrices. The corresponding community detection methods cluster the projected node vectors using the $k$-means clustering method. Each method produces a network partition to represent the community structure. To compare the partition found by these methods with the answer network partition, we adopt the normalized mutual information (NMI) [24] to reflect the effectiveness of each method. The larger the NMI is, the more effective the tested method. As shown in Fig. 4.13, the same to the first test, the normalized Laplacian matrix and the correlation matrix give the best and almost identical results. The adjacency matrix and the modularity matrix also exhibit the similar performance, being a little worse than the normalized Laplacian matrix and the correlation matrix. As to the standard Laplacian matrix with the worst performance, the NMI even reaches 0.4 when the mixing ratio $\mu$ is up to 0.6 with $\gamma = 2$. Furthermore, the heterogeneous distribution of the node degree affects the NMI of the spectral methods based on the adjacency matrix, the modularity matrix and especially the standard Laplacian matrix.

In summary, the normalized Laplacian matrix and the correlation matrix outperforms the other three matrices both at identifying the number of communities according to the spectrum and identifying the community structure using the top eigenvectors. This indicates that it is crucial to take into account the heterogeneous

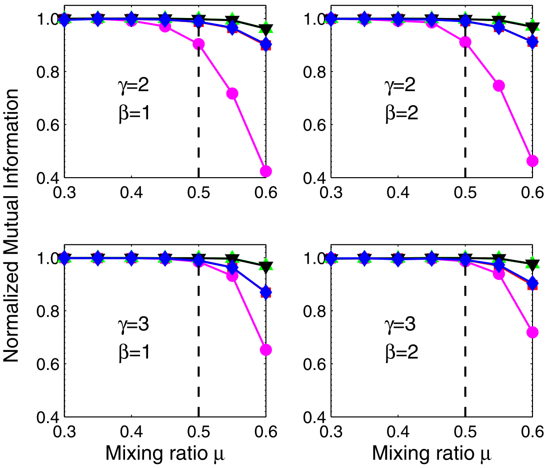


**Fig. 4.13** Effectiveness comparison at identifying intrinsic community structure. The benchmark networks are generated with four different parameter configurations. Each point corresponds to an average over 100 network realizations for each parameter configuration. The corresponding matrices are respectively the adjacency matrix (□), the standard Laplacian matrix (○), the normalized Laplacian matrix (△), the modularity matrix (◊) and the correlation matrix (▽). Reprinted from Ref. [34], Copyright 2010, with permission from IOP Publishing and SISSA

distribution of node degree when using spectral analysis for the detection of community structure. In addition, although the modularity considers the heterogeneity through introducing the null-model reference network (i.e., the configuration model), as shown in [32], this operation is in fact a kind of translation transformation and thus cannot alleviate the influence on the detection of community structure from the heterogeneous distribution of node degree. This phenomenon can be seen from the experimental results on the Lancichinetti's benchmark networks, i.e., the modularity matrix obtains very similar results to the adjacency matrix.

## 4.4 Conclusions

In this chapter, we have studied the diffusion dynamics on networks and the detection of community structure associated with network dynamics.

By analyzing the transients in diffusion process occurring on networks, we find that several stable local equilibrium states emerge during the diffusion process on networks with community structure. These stable states reveal the intrinsic community structure of the underlying networks. Furthermore, as pointed out in this chapter, the spectrum of normalized Laplacian matrix provides critical indicators for the detection of community structure associated with diffusion dynamics on networks. Based on this finding, we proposed a conductance optimization method to identify the community structure, which naturally reflects the diffusion capability of the network. This provides new insights into the number of communities and the multiple topological scales of complex network.

Besides the normalized Laplacian matrix, we conduct a comparative analysis of five matrices on the benchmark networks with heterogeneous distributions of node degree and community size. This comparison is fair and meaningful since the performance of spectral methods heavily relies on the characteristics of the underlying matrices. The comparison is carried out from two perspectives. The former one focuses on whether the number of intrinsic communities can be exactly identified according to the spectrum of these five matrices. The latter evaluates the effectiveness of these matrices at identifying the intrinsic community structure using their eigenvectors. Test results show that the normalized Laplacian matrix and the correlation matrix significantly outperform the other three matrices. The possible reason is that these two matrices are both normalized using the degree of nodes. Thus we can conclude that it is crucial to take into account the heterogeneous distribution of node degree when using spectral analysis for the detection of community structure. In addition, to our surprise, the modularity matrix exhibits very similar performance to the adjacency matrix, which indicates that the modularity matrix does not gain desired benefits from using the configuration model as reference network with the consideration of the node degree heterogeneity.

## References

1. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**, 7821–7826 (2002)
2. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
3. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)
4. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA **103**, 8577–8582 (2006)
5. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA **105**, 1118–1123 (2008)
6. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. P10008 (2008)
7. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
8. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Phys. Rep. **424**, 175–308 (2006)
9. Arenas, A., Dìaz-Guilera, A., Kurths, J., Moreno, Y., Zhou, C.: Synchronization in complex networks. Phys. Rep. **469**, 93–153 (2008)
10. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proc. Natl. Acad. Sci. USA **104**, 36–41 (2007)
11. Arenas, A., Fernández, A., Gómez, S.: Analysis of the structure of complex networks at different resolution levels. New J. Phys. **10**, 053039 (2008)
12. Arenas, A., Díaz-Guilera, A., Pérez-Vicente, C.J.: Synchronization reveals topological scales in complex networks. Phys. Rev. Lett. **96**, 114102 (2006)
13. Pons, P., Latapy, M.: Computing communities in large networks using random walks. Lect. Notes Comput. Sci. **3733**, 284–293 (2005)

14. Zhou, H.: Network landscape from a Brownian particle's perspective. Phys. Rev. E **67**, 041908 (2003)
15. Zhou, H.: Distance, dissimilarity index, and network community structure. Phys. Rev. E **67**, 061901 (2003)
16. Delvene, J.C., Yaliraki, S.N., Barahona, M.: Stability of graph communities across time scales. arXiv:0812.1811
17. Noh, J.D., Rieger, H.: Random walks on complex networks. Phys. Rev. Lett. **92**, 118701 (2004)
18. Barrat, A., Barthélemy, M., Vespignani, A.: Dynamical Processes on Complex Networks. Cambridge University Press, Cambridge (2008)
19. Cheng, X.Q., Shen, H.W.: Uncovering the community structure associated with the diffusion dynamics on networks. J. Stat. Mech. P04024 (2010)
20. Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**, 395–416 (2008)
21. Yu, S.X., Shi, J.: In: Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 313–319 (2003)
22. Evans, T.S., Lambiotte, R.: Line graphs, link partitions, and overlapping communities. Phys. Rev. E **80**, 016105 (2009)
23. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**, 046110 (2008)
24. Danon, L., Duch, J., Diaz-Guilera, A., Arenas, A.: Comparing community structure identification. J. Stat. Mech. P09008 (2005)
25. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**, 452–473 (1977)
26. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait? Behav. Ecol. Sociobiol. **54**, 396–405 (2003)
27. Chauhan, S., Girvan, M., Ott, E.: Spectral properties of networks with community structure. Phys. Rev. E **80**, 056114 (2009)
28. Fiedler, M.: Property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czechoslov. Math. J. **25**, 619–633 (1975)
29. Yan, G., Chen, G.R., Lü, J.H., Fu, Z.Q.: Synchronization performance of complex oscillator networks. Phys. Rev. E **80**, 056116 (2009)
30. Capocci, C., Servedio, V.D.P., Caldarelli, G., Colaiori, F.: Detecting communities in large networks. Physica A **352**, 669–676 (2005)
31. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**, 036104 (2006)
32. Shen, H.W., Cheng, X.Q., Fang, B.X.: Covariance, correlation matrix, and the multiscale community structure of networks. Phys. Rev. E **82**, 016114 (2010)
33. Ding, C., He, X.: In: Proceedings of the 21st International Conference on Machine Learning, pp. 225–232 (2004)
34. Shen, H.W., Cheng, X.Q.: Spectral methods for the detection of network community structure: a comparative analysis. J. Stat. Mech. P10020 (2010)
35. Lambiotte, R., Delvenne, J.C., Barahona, M.: Laplacian dynamics and multiscale modular structure in networks. arXiv:0812.1770 (2008)
36. Ronhovde, P., Nussinov, Z.: Multiresolution community detection for megascale networks by information-based replica correlations. Phys. Rev. E **80**, 016109 (2009)
37. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**, 761–764 (2010)
38. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA **101**, 2658–2663 (2004)

# Chapter 5
# Exploratory Analysis of the Structural Regularities in Networks

## 5.1 Introduction

Network provides a powerful tool for representing the structure of complex systems. These networks include social networks [1, 2], information networks [3, 4], and biological networks [1, 5]. Much of recent research on networks actually aims to understand the structural regularities and further to reveal the relationship between such structural regularities and the function of networks [2, 6]. For example, as a widely-studied structural characteristic of network, community structure is of high interest because communities often correspond to functional units such as pathways for metabolic networks and collections of pages on a similar topic on the Web.

Community structure is a kind of assortative structure, in which nodes are divided into groups such that the members within each group are mostly connected with each other [7]. Contrary to community structure, multipartite structure is another important kind of structural regularities observed in real world networks [8]. Multipartite structure means that nodes of network can be divided into groups such that most of edges are across different groups. Beside these salient structural characteristics, other types of structure are also observed in real world networks, such as hierarchical structure and core-periphery structure.

However, existing methods mostly presume that certain type of structure exists in the target network and then devote to detect such structure. This raises concerns to the reliability of the detected structure. On one hand, the assumed structure may not match the intrinsic structure of the target network and thus these methods are not applicable to these situations. On the other hand, several real world networks contain multiple types of structure simultaneously. Most existing methods are designed for certain type of structure and thus cannot detect the broad types of structure. In addition, several unknown types of structure may also exist in networks and a desired method should be able to detect such structure as well. Thus, it is the time to explore multiple types of structural regularities in networks.

In this chapter, we will devote to the exploratory analysis of structural regularities in networks. We first study the networks with only positive links using a blockmodel. Then, we move to explore the signed networks, i.e., networks with both positive and

H.-W. Shen, *Community Structure of Complex Networks*, Springer Theses,
DOI 10.1007/978-3-642-31821-4_5, © Springer-Verlag Berlin Heidelberg 2013

negative links. To deal with sign of links, we extend Newman's mixture model by separately modeling positive links and negative links.

## 5.2 Regularity Exploration in Networks with Only Positive Links

In this section, we focus on exploring the intrinsic structural regularities in network by dividing network nodes into groups such that the members of each group have similar patterns of connections to other groups. A general stochastic blockmodel is proposed to model the network structure. In this model, node groups are represented by unobserved or hidden quantities and the relationships among groups are explicitly modeled by a block matrix as the traditional blockmodels. Then, using the expectation-maximization algorithm, we fit the model to specific network data and detect intrinsic structural regularities of the network without prior knowledge of the type of regularity existing in the network. Compared with existing models, the most prominent strength of our model is the high flexibility. This strength enables it to possess the advantages of existing models and to overcome their shortcomings in a unified way. As a result, not only broad types of structure can be detected, but also the type of identified structure can be indicated by the block matrix. In addition, our model can tell us the centrality of the node in each group and the mixed membership of nodes as well.

Tests on a number of artificial and real world networks demonstrate that our model outperforms the state-of-the-art models at shedding light on the structural regularities of network, including the overlapping community structure, multipartite structure and several other types of structure which are beyond the capability of existing models.

### *5.2.1 Background*

Recently, several probabilistic generative models are proposed to model network data and to explore the structural regularities [9, 10]. These models view network structure as observed quantities and take communities as hidden groups of nodes. The communities are then identified by fitting the model to the observed network structure. For example, Ren et al. [11] proposed a probabilistic model to uncover the overlapping community structure. This model assumes that the two end nodes of each edge are from the same community and this assumption is satisfied by the fuzzy membership of nodes. Zhang et al. [12] applied the Latent Dirichlet Allocation (LDA, a well-known generative model) to social network analysis and gave a method to detect community structure. The common drawback of these two models is that they can only uncover the community structure and fail to reveal other types of structural regularities, e.g., multipartite structure.

To characterize the hierarchical organization of networks, Clauset et al. proposed the hierarchical random graph model, which is capable of expressing both assortative and disassortative structure [13]. To explore more broad types of structure, Newman et al. proposed a mixture model for exploratory analysis of network structure [14]. In this model, the nodes with similar connection preference rather than the highly connected nodes are classified into the same group. In such a general way, this model can reveal several other kinds of structural regularities beyond community structure. However, this model fails to tell us which kind of structural regularities has been identified. More importantly, this model may produce a result which is a mixture of several types of structure, and thus the identified structure may not provide clear information about the structural regularities. The shortcoming of this model is attributed to that it only models the relationship between groups and nodes rather than the relationship among groups. Stochastic blockmodel provides an appropriate alternative to the mixture model for exploring broad range of structural regularities. Karrer et al. utilized a degree-corrected stochastic blockmodel [15] to investigate community structure of network. Airoldi et al. gave a mixed membership stochastic blockmodel [16] to model network data. These works have demonstrated that stochastic blockmodel is a good choice for exploring regularities of network. However, the effectiveness of these models is limited by their inflexible model assumptions, e.g., the hard partition assumption or neglecting the directionality of edges.

### *5.2.2 The Stochastic Blockmodel*

Generally, a network with $n$ nodes can be represented mathematically by an adjacency matrix $A$ with elements $A_{ij} = 1$ if there is an edge from node $i$ to node $j$ and 0 otherwise. For weighted networks, $A_{ij}$ is generalized to represent the weight of the edge from $i$ to $j$.

To investigate the structural regularities in network, we suppose that the $n$ nodes of the network fall into $c$ groups whose memberships are unknown, i.e., we cannot observe or measure them directly. Here, we propose a statistical model to infer the group membership from the observed network structure.

The model we used is a kind of stochastic blockmodel. Blockmodel is a generative model and has a long tradition of study in the social science and computer science. For a standard blockmodel, a $c \times c$ matrix $\omega$ is generally adopted such that the matrix element $\omega_{rs}$ denotes the probability that a randomly selected edge connects group $r$ to group $s$, i.e., the tail node of the edge is from group $r$ and the head node is from $s$. The advantage of blockmodel lies in that the matrix $\omega$ explicitly characterizes various types of connecting patterns among groups.

In the standard blockmodel, the nodes in the same group are identical, i.e., each node in a group has equal probability to be the end node of an edge adjacent to the group. This constraint is relaxed in our model. Specifically, for an edge with its tail node being from group $r$ and its head node being from group $s$, we use $\theta_{ri}$ to denote the probability that the tail node is $i$ and $\phi_{sj}$ to denote the probability that the head node is $j$ respectively. In addition, we use $\overrightarrow{g}_{ij}$ and $\overleftarrow{g}_{ij}$ to denote respectively the group membership of the tail node and head node of the edge $e_{ij}$.

Up to now, we have given all the quantities in our model. They can be classified into three classes: observed quantities $\{A_{ij}\}$, hidden quantities $\{\overrightarrow{g}_{ij}, \overleftarrow{g}_{ij}\}$, and model parameters $\{\omega_{rs}, \theta_{ri}, \phi_{sj}\}$. To simplify the notations, we henceforth denote by $A$ the entire set $\{A_{ij}\}$ and similarly $\overrightarrow{g}$, $\overleftarrow{g}$, $\omega$, $\theta$, $\phi$ for $\{\overrightarrow{g}_{ij}\}$, $\{\overleftarrow{g}_{ij}\}$, $\{\omega_{rs}\}$, $\{\theta_{ri}\}$ and $\{\phi_{sj}\}$.

With our model, an edge $e_{ij}$ is generated in the following process:

1. Select two groups $\overrightarrow{g}_{ij} = r$ and $\overleftarrow{g}_{ij} = s$ respectively for the tail node and head node of the edge with probability $\omega_{rs}$;
2. Draw the tail node $i$ from the group $r$ with probability $\theta_{ri}$;
3. Draw the head node $j$ from the group $s$ with probability $\phi_{sj}$.

Summing over the latent quantities $r$ and $s$, the probability that we observe an edge $e_{ij}$ can be written as

$$\Pr(e_{ij}|\omega, \theta, \phi) = \sum_{rs} \omega_{rs}\theta_{ri}\phi_{sj}. \tag{5.1}$$

Then, the likelihood of the observed network with respect to our model is

$$\Pr(A|\omega, \theta, \phi) = \prod_{ij}\left(\sum_{rs} \omega_{rs}\theta_{ri}\phi_{sj}\right)^{A_{ij}}. \tag{5.2}$$

Note that the self-loop edges are allowed and the weight $A_{ij}$ is taken as the number of multi-edges connecting node $i$ to node $j$ as done in many existing models including, for instance, the widely studied configuration model [17].

Intuitively, the parameter $\theta_{ri}$ characterizes the centrality of node $i$ in the group $r$ from the perspective of outgoing edges while $\phi_{sj}$ describes the centrality of node $j$ in the group $s$ from the perspective of incoming edges. Differently from traditional blockmodels, by differentiating these two kinds of centrality, our model can provide more flexibility to explore broad types of intrinsic structural regularities in network. Note that the parameters $\omega_{rs}$, $\theta_{ri}$, $\phi_{sj}$ satisfy the normalization conditions

$$\sum_{r=1}^{c}\sum_{s=1}^{c} \omega_{rs} = 1, \qquad \sum_{i=1}^{n} \theta_{ri} = 1, \qquad \sum_{j=1}^{n} \phi_{sj} = 1. \tag{5.3}$$

Now our task is to estimate the model parameters and to infer the unobserved quantities by fitting the model to the observed network data. The standard framework for such a task is likelihood maximization. Generally, one works not with the likelihood [Eq. 5.2] itself but with its logarithm (log-likelihood)

$$\begin{aligned}\mathscr{L} &= \ln \Pr(A|\omega, \theta, \phi) \\ &= \sum_{ij} A_{ij} \ln\left(\sum_{rs} \omega_{r,s}\theta_{ri}\phi_{sj}\right). \end{aligned} \tag{5.4}$$

The maximum of the likelihood and its logarithm are in the same place since the logarithm is a monotonically increasing function.

Directly maximizing the log-likelihood is difficult because of the inner sum over the unobserved quantities $\overrightarrow{g}_{ij} = r$ and $\overleftarrow{g}_{ij} = s$. Using Jensen's inequality, the maximization of the log-likelihood can be transformed into the maximization of the expected log-likelihood

$$\begin{aligned}\overline{\mathscr{L}} &= \sum_{\overrightarrow{g}, \overleftarrow{g}} \Pr(\overrightarrow{g}, \overleftarrow{g} | A, \omega, \theta, \phi) \ln \Pr(A | \overrightarrow{g}, \overleftarrow{g}, \omega, \theta, \phi) \\ &= \sum_{ijrs} \Pr(\overrightarrow{g}_{ij} = r, \overleftarrow{g}_{ij} = s | e_{ij}, \omega, \theta, \phi) \big[ A_{ij} (\ln \omega_{rs} + \ln \theta_{ri} + \ln \phi_{sj}) \big] \\ &= \sum_{ijrs} q_{ijrs} A_{ij} (\ln \omega_{rs} + \ln \theta_{ri} + \ln \phi_{sj}), \end{aligned} \tag{5.5}$$

where to simplify the notation we have defined $q_{ijrs} = \Pr(\overrightarrow{g}_{ij} = r, \overleftarrow{g}_{ij} = s | e_{ij}, \omega, \theta, \phi)$, which denotes the probability that one observes an edge $e_{ij}$ with its tail node $i$ from group $r$ and its head node $j$ from group $s$ given the observed network and the model parameters.

With the expected log-likelihood, we can give the best estimate of the value $\overline{\mathscr{L}}$ and the position of its maximum represents the best estimate of the most likely values of the model parameters. Specifically, if the value of $q_{ijrs}$ is known, we can find the values of the model parameters $\omega$, $\theta$, $\phi$ where $\overline{\mathscr{L}}$ reaches its maximum. However, the calculation of $q_{ijrs}$ requires the values of these model parameters. To address such a problem, an expectation-maximization (EM) algorithm is adopted.

Under the framework of EM algorithm, we first calculate the value of $q_{ijrs}$ by

$$\begin{aligned} q_{ijrs} &= \frac{\Pr(\overrightarrow{g}_{ij} = r, \overleftarrow{g}_{ij} = s, e_{ij} | \omega, \theta, \phi)}{\Pr(e_{ij} | \omega, \theta, \phi)} \\ &= \frac{\omega_{rs} \theta_{ri} \phi_{sj}}{\sum_{rs} \omega_{rs} \theta_{ri} \phi_{sj}}. \end{aligned} \tag{5.6}$$

Once we have the values of the $q_{ijrs}$, we can use them to evaluate the expected log-likelihood and hence to find the values of $\omega$, $\theta$, $\phi$ that maximize it.

Introducing the Lagrange multipliers $\rho$, $\gamma_r$ and $\lambda_s$ to incorporate the normalization conditions in Eq. 5.3, the expected log-likelihood expression to be maximized becomes

$$\widetilde{\mathscr{L}} = \overline{\mathscr{L}} + \rho \left(1 - \sum_{rs} \omega_{rs}\right) + \sum_r \gamma_r \left(1 - \sum_i \theta_{ri}\right) + \sum_s \lambda_s \left(1 - \sum_j \phi_{sj}\right). \tag{5.7}$$

By letting the derivative of $\widetilde{\mathscr{L}}$ to be 0, the maximum of the expected log-likelihood occurs at the places where

$$\begin{cases} \omega_{rs} = \dfrac{\sum_{ij} A_{ij} q_{ijrs}}{\sum_{ijrs} A_{ij} q_{ijrs}}, \\ \theta_{ri} = \dfrac{\sum_{js} A_{ij} q_{ijrs}}{\sum_{ijs} A_{ij} q_{ijrs}}, \\ \phi_{sj} = \dfrac{\sum_{ir} A_{ij} q_{ijrs}}{\sum_{ijr} A_{ij} q_{ijrs}}. \end{cases} \tag{5.8}$$

Equations 5.6 and 5.8 constitute our expectation-maximization algorithm. In the expectation step, the expected value of log-likelihood is calculated through evaluating the values of $q_{ijrs}$ with Eq. 5.6. In the maximization step, the expected value of log-likelihood is maximized when the values of model parameters $\omega$, $\theta$, $\phi$ are evaluated with Eq. 5.8. Implementation of the algorithm consists merely of iterating Eqs. 5.6 and 5.8 until convergence.

When the algorithm converges, we obtain a set of values for hidden quantity $q_{ijrs}$ and model parameters $\omega$, $\theta$, $\phi$. This set of values is self-consistent with respect to Eqs. 5.6 and 5.8. However, it is not always the place where the log-likelihood reaches its maximum. In other words, the expectation-maximization algorithm may converge to local maxima of the log-likelihood. With different starting values, the algorithm will give rise to different solutions. To obtain a satisfactory solution, it is necessary to perform many runs with different starting values of model parameters and take the solution giving the highest log-likelihood over all the runs performed.

By fitting the model to the observed network structure with the expectation-maximization algorithm, the estimated model parameters provide us vital information for structural regularities of the network. Specifically, $\theta$ and $\phi$ describe the centrality of a node in groups containing it respectively from the perspective of outgoing edges and incoming edges. The parameter $\omega$ characterizes the connecting patterns among different groups, i.e., the type of structural regularities.

More importantly, according to the model parameters, we can define two kinds of group memberships $\alpha_{ir}$ and $\beta_{js}$ respectively from the perspective of outgoing edges and incoming edges. Specifically, $\alpha_{ir}$ is the probability that node $i$ is from group $r$ when it acts as the tail node of edges while $\beta_{js}$ is the probability that node $j$ is from group $s$ when it acts as the head node of edges. For $\alpha_{ir}$, it can be calculated by

$$\alpha_{ir} = \frac{\sum_s \omega_{rs}\theta_{ri}}{\sum_{rs} \omega_{rs}\theta_{ri}}. \tag{5.9}$$

Actually, $\alpha_{ir}$ provides a soft or fuzzy membership, i.e., node $i$ can belong to more than one group simultaneously. When the identified structural regularity corresponds to community structure, we actually obtain the overlapping community structure which has attracted much research attention ever since it is proposed. If one wants to get a hard partition, we can simply assign each node $i$ to the group $r$ satisfying $r = \arg\max_s\{\alpha_{is}, s = 1, 2, \ldots, c\}$. These statements for $\alpha_{ri}$ also apply to $\beta_{ir}$ defined as

$$\beta_{js} = \frac{\sum_r \omega_{rs}\phi_{sj}}{\sum_{rs} \omega_{rs}\phi_{sj}}. \tag{5.10}$$

Finally, the model described above so far is based on directed networks. Actually, the model can be easily generalized to undirected networks by letting the parameter $\theta$ be identical to $\phi$. The derivation follows the case of directed networks and the results are the same to Eqs. 5.6 and 5.8.

Now we discuss the computational cost of the expectation-maximization algorithm for the fitting of our model. For each iteration in this algorithm, the cost consists of two parts. The first part is from the calculation of $q_{ijrs}$ using Eq. 5.6, whose time-complexity is $O(m \times c^2)$. Here $m$ is the number of edges in the network and $c$ is the number of groups. The second part is from the estimation of the model parameters using Eq. 5.8, whose time-complexity is also $O(m \times c^2)$. We use $T$ to denote the average number of iterations before the iteration process converges. Then, the total cost of the expectation-maximization algorithm for our model is $O(K \times T \times m \times c^2)$. Here, $K$ is the number of times that the iteration process is restarted with different starting values to obtain a satisfactory solution. It is difficult to give a theoretical estimation to the number $T$ of iterations. Generally speaking, $T$ is determined by the network structure and the starting values of model parameters. The number of runs is dependent on the scale of the network and its structural characteristics. For the networks tested here, only less than 10 runs are enough to obtain a satisfactory result.

The computational cost limits our model to dealing with networks with tens of thousands of nodes. We look forward to seeing more efficient implementation for our model. Note that the method proposed in [18] provides a promising way to improve the computational efficiency and to decrease the memory space required. Finally, for the convenience to evaluate the results and to apply our model to more real world networks, we make the source computer code of our model available.

### *5.2.3 Comparison with Other Models*

In this section, we illustrate the difference and connections between our model with several existing models. Figure 5.1 gives the schematic for our model and two existing generative models, namely Newman's mixture model and Ren's probabilistic model.

For Newman's model, as shown in Fig. 5.1(a), each group $r$ is characterized by the connecting preference $\theta_{rj}$ to node $j$, no matter the node $j$ is contained by the group $r$ or not. The nodes belonging to the same group have similar connecting preference. As a result, both assortative and disassortative structural regularities can be detected by this model. However, this model has no parameter to explicitly characterize the type of the identified structure. More importantly, this model may produce a result which is a mixture of several types of structure and thus in these cases the identified structure may bring confused information about the structural regularities. For example, for the karate club network [20] shown in Fig. 5.2, nodes 12, 15, 16, 19, 21, 23 are identified by this model as overlapped nodes shared by the two groups, denoted by circles and squares, although these nodes only have connections to one of the two groups.

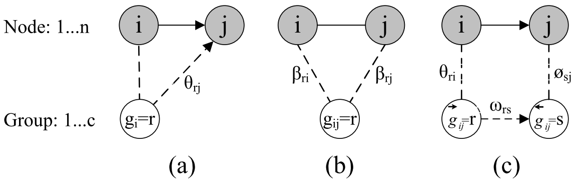


**Fig. 5.1** Illustration of three generative models for network data (**a**) Newman's mixture model, (**b**) Ren's probabilistic model, and (**c**) our model. *Filled circles* represent observed quantities and unfilled ones correspond to hidden quantities. The *solid line* (with arrow) between node $i$ and $j$ indicates the existence of one (directed) edge connecting them. The *dashed-line* connecting two circles indicates that the relation between the corresponding quantities is unobserved and requires being learned from the observed network data. *Arrows* represent the directions of relation. Reprinted with permission from Ref. [19]. Copyright 2011 by American Physical Society

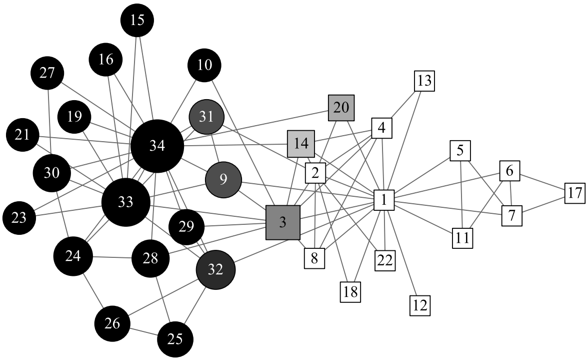


**Fig. 5.2** Results on the karate club network. The real social fission of this network is represented by two different shapes, circle and square. The shades of nodes indicate the mixed membership obtained by fitting our model to this network. The sizes of the nodes indicate the centrality degree (i.e., $\theta_{ri}$) of nodes with respect to the left group. Here, $\theta_{ri}$ ranges from 0 for the smallest nodes to 0.22 for the largest. Reprinted with permission from Ref. [19]. Copyright 2011 by American Physical Society

For Ren's model, as shown in Fig. 5.1(b), the two end nodes of each edge are assumed to be from the same group. As a result, only the assortative structure (community structure) can be detected using this model. Note that, for this model, no edge is allowed to connect different groups. The relationship between communities is reflected by the overlapped nodes.

For our model, it essentially is a kind of stochastic blockmodel, in which the relationships among different node groups are explicitly modeled by the block matrix $w$. In this way, our model possesses the advantages of both Newman's model and Ren's model and overcomes the shortcoming of these two models.

On one hand, through learning the matrix $w$ according to observed network data, various types of structural regularities can be explored by our model. The type of the identified structure is indicated by the matrix $w$. Specifically, when the matrix $\omega$ is an identity matrix, the identified structural regularity corresponds to an obvious community structure. Meanwhile, multipartite or anti-community structure is revealed when the estimated model parameter $\omega$ is an anti-diagonal matrix with all the anti-diagonal elements being 1. For other types of structure such as core-periphery structure and hierarchical structure, the form of $\omega$ is the same to the block matrix $\omega$ in traditional block models [15].

On the other hand, using the matrix $w$, our model discards the assumption of Ren's model that two end nodes of one edge are required to be from the same community. In this sense, Ren's model is a special case of our model. In addition, our model also provides several other flexibilities. By representing the centrality of nodes in group from two different perspectives respectively according to the outgoing edges and incoming edges, our model can detect more broad range of structural regularities which is out of the capability of other models. This will be shown later in the subsequent section. Moreover, our model can be further generalized by not requiring the matrix $w$ be a square matrix.

Finally, we compare our model to two recently proposed stochastic models for community detection [15, 18]. Firstly, both our model and Karrer's model [15] are stochastic blockmodel where a block matrix is adopted to characterize the connecting patterns among groups. The main difference between these two models lies in that Karrer's model is designed to detect disjoint structural regularities while our model is for fuzzy structural regularities. This difference is reflected by the definition of the model parameters $\theta$ and $\phi$ in our model and the definition of the model parameter $\theta$ in Karrer's model. In addition, our model differentiates the outgoing edges from incoming edges of nodes while Karrer's model does not. Secondly, similar to Ren's model, Ball's model [18] focused on the community structure while our model can uncover multiple types of structural regularities.

## 5.2.4 Experimental Results

In this section, we demonstrate the effectiveness of our model at exploring the structural regularities of networks by experiments on several real world or artificial networks with various types of intrinsic structural regularities. Then we discuss the model selection issue, i.e., how to determine the optimal number of groups.

### 5.2.4.1 Detecting Community Structure

The test network is the famous karate club network constructed by Zachary [20]. This network characterizes the acquaintance relationship between 34 members of a karate club in an American University. A dispute arose between the club's administrator and its principal karate teacher, and as a result the club eventually split into two smaller clubs, centered around the administrator and the teacher respectively. The network and its fission are depicted in Fig. 5.2. The administrator and the teacher are represented by nodes 1 and 33 respectively.

**Table 5.1** Mixed membership of overlapped nodes. Reprinted with permission from Ref. [19]. Copyright 2011 by American Physical Society

| Node ID | $\alpha_{i1}$ | $q_{i1}$ [a] | $\frac{u_{1i}}{u_{1i}+u_{2i}}$ [b] |
|---|---|---|---|
| 3 | 0.49 | 0.00 | 0.49 |
| 9 | 0.70 | 0.96 | 0.70 |
| 14 | 0.24 | 0.00 | 0.24 |
| 20 | 0.33 | 0.13 | 0.33 |
| 31 | 0.71 | 0.92 | 0.71 |
| 32 | 0.83 | 1.00 | 0.83 |

[a] $q_{i1}$ is defined in [14] as the probability that node $i$ belongs to group 1

[b] $\frac{u_{1i}}{u_{1i}+u_{2i}}$ is defined in [11] as the probability that node $i$ belongs to group 1

By setting the group number $c = 2$, we fit our model to the karate club network data. The resulted matrix $\omega$ is a 2 × 2 identity matrix, indicating that the obtained structure is community structure. Figure 5.2 shows the two groups found by our model with the expectation-maximization method. As shown in Fig. 5.2, the shades of the nodes in the figure represent the values of $\alpha_{i1}$,[1] where group 1 is the left group. As we can see, our model assigns most of the nodes strongly to one group or the other. Actually, all but 6 nodes are assigned 100 % to one of the groups (black and white nodes in the figure). If we simply divide the nodes into two disjoint groups by assigning each node $i$ to the group $r$ according to the belong coefficients $\alpha_{ir}$, the resulting groups perfectly correspond to the real split of the club.

In addition, Table 5.1 gives the belonging coefficient of the 6 overlapped nodes which are shared by the two groups. These overlapped nodes are nodes 3, 9, 14, 20, 31, 32. Note that these overlapped nodes are often misclassified by traditional partition-based community detection methods. For comparison, we also give the mixed membership of these six nodes according to Newman's mixture model and Ren's model. As we can see, our model and Ren's model produce the same results, which is attributed to the fact that Ren's model is a special case of our model. However, Newman's model behaves very differently from the other two models. Actually, for Newman's model, another 10 nodes are also assigned to all the two groups, e.g., nodes 12, 15. Such a result is counterintuitive to the real structure of this network. As a conclusion, our model performs better than Newman's model at detecting the overlaps between groups. Ren's model can only detect community structure while our model can detect other types of structural regularities as illustrated in the following test.

#### 5.2.4.2 Detecting Multipartite Structure

Now we illustrate the detection of multipartite or anti-community structure according to our model. The test network is the adjacency network of English words taken

[1] Since this network is an undirected network, the two kinds of belonging coefficient are identical, i.e., $\alpha_{ir} = \beta_{ir}$.

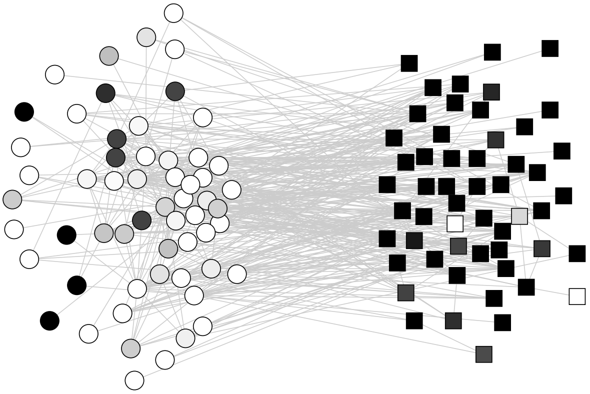


**Fig. 5.3** The adjacency network of English words. Node groups corresponding to adjectives and nouns are respectively denoted by circle and square. The shades of nodes indicate their belonging coefficient obtained by fitting our model to this network. Reprinted with permission from Ref. [19]. Copyright 2011 by American Physical Society

from Ref. [8]. In this network, the nodes represent 112 commonly occurring adjectives and nouns in the novel *David Copperfield* by Charles Dickens, with edges connecting any pair of words that appear adjacent to each other at any place in the text. Generally, adjectives occur next to nouns in English. Thus most edges in the network connect an adjective to a noun and the network is approximately bipartite, i.e., this network possesses anti-community structure. This can be seen clearly in Fig. 5.3, where the adjectives and nouns are respectively represented by circles and squares.

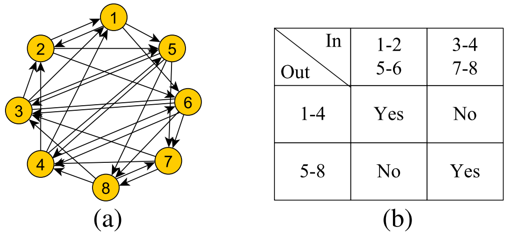
Fitting our model to this network with $c = 2$, the resulted $\omega$ is a $2 \times 2$ anti-diagonal matrix, indicating that the identified structure is bipartite structure. The obtained two groups and node memberships are shown by the shades of nodes as shown in Fig. 5.3. We can see that most nodes are assigned to only one group, although there are several ambiguous cases corresponding to the nodes with intermediate shades. If we assign each node to its most preferred group, the resulted two disjoint groups well separate the adjectives from the nouns. In fact, 100 of all the 112 nodes are correctly classified. This accuracy is the same to the result given by Newman's mixture model.

As a comparison, we also apply Ren's model to this network by setting the group number being 2. Only 60 nodes of all the 112 nodes are correctly classified, similar to the accuracy of random assignment. The ineffectiveness of Ren's model at this network is attributed to that Ren's model presumes the existence of community structure in the network while the intrinsic structural regularity is bipartite structure.

### 5.2.4.3 Exploring Other Type of Structural Regularity

In the previous tests, we have demonstrated that our model can be used to detect both the assortative structure (i.e., community structure) and the disassortative structure (i.e., multipartite structure) without being told that which type of structural regularities exists in the target networks. Now we will further show that our model can also detect other types of structure which cannot be revealed by competing models.

We consider the schematic network depicted in Fig. 5.4(a). This network is constructed according to the rules in Fig. 5.4(b). Intuitively, according to the outgoing



| In<br>Out | 1-2<br>5-6 | 3-4<br>7-8 |
|---|---|---|
| 1-4 | Yes | No |
| 5-8 | No | Yes |

**Fig. 5.4** A schematic network. The directed edges are placed according to the rules described the *right table*. Reprinted with permission from Ref. [19]. Copyright 2011 by American Physical Society

edges in this network, the nodes can be divided into two groups: $\{1, 2, 3, 4\}$ and $\{5, 6, 7, 8\}$. Meanwhile, according to the incoming edges, the nodes of this network belong to another two groups: $\{1, 2, 5, 6\}$ and $\{3, 4, 7, 8\}$.

We apply Newman's model, Ren's model and our model to this schematic network. Limited by the assumptions of models, both Newman's model and Ren's model fail to uncover the intrinsic structural regularity indicated by the construction rules. For our model, the flexibility of model assumption enables it to accurately detect such type of structure. Specifically, by fitting our model to this network, the obtained $\theta$ or $\alpha$ reveals the two groups indicated by the outgoing edges while the $\phi$ or $\beta$ reflects the two groups indicated by the incoming edges.

#### 5.2.4.4 Model Selection Issue

In the previous tests, we need to specify the group number before fitting our model to network. However, the group number is unknown a prior for many cases. Thus it is helpful to give a criterion to determinate the appropriate group number for given network. This task is known as the model selection issue in statistics. We deal with this problem by using minimum description length principle, which is also used to handle the model selection issue in Ren's model.

According to minimum description length principle, the required length to describe the network data is composed of two parts. The first part describes the coding length of the network using our model. This coding length is $-L$ for directed network and $-L/2$ for undirected network. The second part gives the length for coding model parameters. This part is $-\sum_{rs} \ln \omega_{rs} - \sum_{ri} (\ln \theta_{ri} + \ln \phi_{ri})$ for directed network and $-\sum_{rs} \ln \omega_{rs} - \sum_{ri} \ln \theta_{ri}$ for undirected network. In this way, the optimal $c$ is the one which minimizes the total description length.

As tests, we consider two real world networks with prior knowledge of the intrinsic group numbers. These two networks are respectively the journal citation network constructed in Ref. [21] and the American football team network described in Ref. [1]. In the journal citation network, each node corresponds to a journal and all the 40 journals are from four different fields: multidisciplinary physics, chemistry, biology and ecology. Journals from the same field are more likely connected by citation relation. For the football network, nodes represent the 115 teams respectively belonging to 12 conferences and generally games are more frequent between members of the same conference than between teams of different conferences.

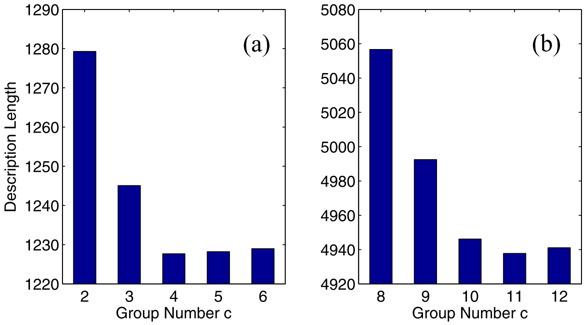


**Fig. 5.5** Model selection results. (**a**) Journal citation network and (**b**) American football team network. Reprinted with permission from Ref. [19]. Copyright 2011 by American Physical Society

As shown in Fig. 5.5, the number of intrinsic groups is correctly identified for the journal citation network. However, for the football network, 11 is the optimal number of groups while the intrinsic number is 12. By checking the found node groups, we find that only 11 node groups have their identities, i.e., each group contains at least one node after assigning nodes to their most preferred groups according to the obtained belonging coefficient $\alpha$ or $\beta$. This indicates that the appropriate group number is 11 for the football network. In fact, many well-known community detection methods also identify 11 communities.

### *5.2.5 Summary*

In this section, we have studied the exploration of intrinsic structural regularities in network using a general stochastic blockmodel. Without prior knowledge, our model not only can detect broad types of intrinsic structural regularities, but also can learn the type of identified structure directly from the network data. Tests on a number of artificial and real world networks demonstrate that our model outperforms the state-of-the-art models at shedding light on the structural features of networks. The flexibility enables our model to be an effective way to reveal the structural regularities of network and further to help us to understand the relationship between structure and function of network. For potential applications, our model can be used to predict the emergence or vanishing of edges in network. This model can be generalized by releasing the requirement that the block matrix is a square matrix and investigate the possible applications of the more flexible model.

## 5.3 Regularity Exploration in Signed Networks

In this section, we explore the structural regularities in signed networks, i.e., networks with both positive and negative links. Specifically, we generalize Newman's mixture model to explore the broad types of structural regularities in signed network. For Newman's mixture model which is designed to deal with networks with only positive links, the nodes in the same group have similar connecting preference to other nodes, i.e., the nodes in the same group have many common friends. From the perspective of social balance, we generalize such an idea to networks with both positive and negative links. The nodes in the same group either have many common friends or have many common enemies. In this way, we propose a model to explore both the assortative and disassortative structural regularities in networks with both positive and negative links. In addition, compared with existing methods for detecting the community structure in networks containing negative links, our model possesses an distinct advantage, i.e., it can explore both the assortative and the disassortative structural regularities in a unified way from the perspective of social balance. Finally, the effectiveness of our model is demonstrated by tests on a number of real world networks.

### *5.3.1 Background*

These existing methods all assume that the target networks contain only positive links. However, negative links also exist in real world networks. These links may reflect the enmity between individuals or organizations. Typical examples include the unfriendly relations among persons, the competing relations among companies, and the disputes among countries. They may also represent the relations among anti-correlated objects. Such kind of examples includes the deactivating relation among neurons and the repulsions between different kinds of retailer stores. Empirical studies have shown that the coupled relation between positive links and negative links is critical to the evolution of the whole networks [22]. Recently, social balance theory developed by Heider [23] has been enriched to analyze the dynamics of networks with both positive and negative links [24]. More importantly, the conflicts emerged from highly interacting individuals could find its root in the couple between positive and negative links [22]. Therefore, it is crucial to give a method for exploring the structural regularities in networks with both positive and negative links.

Several community detection methods have been extended to networks with positive and negative links. Gomez et al. generalized the widely-used modularity for community detection to deal with negative links [25]. Traag and Bruggeman [26] gave an extended version of Potts Model to detect the community structure in networks with negative links. Rubinov and Sporns [27] gave another variant of modularity to investigate the functional brain networks containing negative links. Yang et al. [28] gave an agent-based method to mine the community structure in signed networks. The common drawback of these methods lies in that they are only capable of detect community structure and fail to uncover other types of structural regularities, e.g., disassortative structure.

## 5.4 Extended Mixture Model for Network Exploration

A directed network is often described by an adjacency matrix $A$. The entries of $A$ are defined as follows: $A_{ij} = 1$ if a positive link is present from node $i$ to node $j$,

$A_{ij} = -1$ if a negative link is present, and $A_{ij} = 0$ otherwise. For weighted networks, $A_{ij}$ is generalized to represent the weight of the link from $i$ to $j$. We further separate the negative links from positive links by setting $A^{+}_{ij} = A_{ij}$ if $A_{ij} > 0$ and 0 otherwise, and $A^{-}_{ij} = -A_{ij}$ if $A_{ij} < 0$ and 0 otherwise. Thus, $A = A^{+} - A^{-}$. For an undirected network, it can be transformed into a directed one simply by replacing each undirected link with two oppositely directed links. Note that, for a self-loop link, only one directed link is used to replace the original undirected one. We hereafter only consider directed networks.

We suppose that the $n$ nodes of network fall into $c$ groups and we denote by $g_i$ the group to which node $i$ belongs. These group memberships are often unknown to us and we cannot measure them directly. For the goal of exploring the structural regularities of network, we need to infer the group memberships from the observed network structure. This is a typical statistical inference problem and the standard solution for such a problem is to give a generative model for the observed network structure and then to determine the parameters of the model by finding the best fit to the observed network.

We generalize Newman's mixture model from networks containing only positive links to networks with both positive and negative links. In the extended model, the probability that we observe the given network is parameterized by three sets of parameters, namely $\pi$, $\theta$ and $\phi$. The parameter $\pi_r$ represents the probability that a randomly chosen node falls into group $r$. Intuitively, it is the fraction of nodes in group $r$. The parameter $\theta_{ri}$ is the probability that a positive link departing from a particular node in group $r$ connects to node $i$. Similarly, the parameter $\phi_{ri}$ is the probability that a negative link from a particular node in group $r$ connects to node $i$. In fact, $\theta_{ri}$ and $\phi_{ri}$ represent the "preferences" of nodes in group $r$ about which other nodes they connect to with positive and negative links respectively. These preferences define a group as a set of nodes that all have similar patterns of connection to other nodes. Note that there is no assumption that the nodes $i$ to which the members of group $r$ connect belong to any particular group, i.e., they can be in the same group or in different groups or randomly distributed over the entire network. This general definition for node groups inherits from Newman's mixture model while the newly added advantage of our model is its capacity to deal with the negative links with the additional model parameter $\phi_{ri}$.

According to social balance theory, such a definition of group is reasonable for networks with both positive and negative links. On one hand, if two nodes both have positive links to other nodes, they are more likely to fall into the same group. This is consistent with the statement in social balance theory that two individuals have many common friends are also friends with high probability. On the other hand, if two nodes both have negative links to other particular nodes, they are also very likely to fall into the same group. This is consistent with the statement that two individuals having many common enemies are friends with high probability. Note that such a definition of group may partly violate the social balance in the sense that three individuals being enemies among any two of them can also belong to the same group. This is attributed to the non-transitivity of negative links.

Note that the parameters $\pi_r, \theta_{ri}, \phi_{ri}$ satisfy the normalization conditions

$$\sum_{r=1}^{c} \pi_r = 1, \qquad \sum_{i=1}^{n} \theta_{ri} = 1, \qquad \sum_{i=1}^{n} \phi_{ri} = 1. \tag{5.11}$$

Up to now, we have introduced all the quantities in our model. They can be classified into three classes: observed quantities $\{A_{ij}\}$, hidden quantities $\{g_i\}$, and model parameters $\{\pi_r, \theta_{ri}, \phi_{ri}\}$. To simplify the notations, we henceforth denote by $A$ the entire set $\{A_{ij}\}$ and similarly $g, \pi, \theta, \phi$ for $\{g_i\}$, $\{\pi_r\}$, $\{\theta_{ri}\}$ and $\{\phi_{ri}\}$.

With our model, a directed network with its adjacency matrix $A_{ij}$ is generated in the following process:

1. For each node $i$, it is assigned to the group $g_i$ with probability $\pi_{g_i}$;
2. For each positive links from node $i$, it connects to node $j$ with probability $\theta_{g_i,j}$;
3. For each negative links from node $i$, it connects to node $j$ with probability $\phi_{g_i,j}$.

Thus, the likelihood $\Pr(A, g|\pi, \theta, \phi)$ can be written as

$$\Pr(A, g|\pi, \theta, \phi) = \prod_i \Big[ \pi_{g_i} \prod_j \theta_{g_i,j}^{A_{ij}^+} \phi_{g_i,j}^{A_{ij}^-} \Big]. \tag{5.12}$$

Note that the self-loop links are allowed and the weight $A_{ij}^+$ and $A_{ij}^-$ are respectively viewed as the number of positive and negative multi-links from node $i$ to node $j$ as done in many existing models including, for instance, the widely studied configuration model [17].

To infer the unobserved group membership $g$, we fit our model to the observed network structure by maximizing the likelihood in Eq. 5.12 with respect to the model parameters $\pi$, $\theta$ and $\phi$. For convenience, one usually works not with the likelihood itself but with its logarithm

$$\begin{aligned} \mathscr{L} &= \ln \Pr(A, g|\pi, \theta, \phi) \\ &= \sum_i \Big[ \ln \pi_{g_i} + \sum_j \big( A_{ij}^+ \ln \theta_{g_i,j} + A_{ij}^- \ln \phi_{g_i,j} \big) \Big]. \end{aligned} \tag{5.13}$$

The maximum of the likelihood and its logarithm occur in the same place since the logarithm is a monotonically increasing function.

Since that the group membership $g$ is unknown in our model, we cannot calculate the value of the log-likelihood $\mathscr{L}$ directly according to Eq. 5.13. However, we can first give a good guess at the value of $g$ given the network structure $A$ and the model parameters $\pi$, $\theta$ and $\phi$, i.e., we can calculate the probability distribution $\Pr(g|A, \pi, \theta, \phi)$. Then using the estimation of $g$, we calculate an expected value for the log-likelihood by averaging over $g$ as follows:

$$\overline{\mathscr{L}} = \sum_g \Pr(g|A, \pi, \theta, \phi) \ln \Pr(A, g|\pi, \theta, \phi)$$

$$= \sum_{ir} \Pr(g_i = r|A, \pi, \theta, \phi)\left[\ln \pi_r + \sum_j \left(A_{ij}^{+} \ln \theta_{rj} + A_{ij}^{-} \ln \phi_{rj}\right)\right]$$

$$= \sum_{ir} q_{ir}\left[\ln \pi_r + \sum_j \left(A_{ij}^{+} \ln \theta_{rj} + A_{ij}^{-} \ln \phi_{rj}\right)\right], \tag{5.14}$$

where to simplify the notation we have defined $q_{ir} = \Pr(g_i = r|A, \pi, \theta, \phi)$, which is the probability that the group membership of node $i$ is $r$ given the observed network structure and the model parameters. Actually, $q_{ir}$ provides vital information on node memberships. Note that $q_{ir}$ satisfies the normalization condition $\sum_r q_{ir} = 1$.

With the expected log-likelihood, we can give the best estimate of the value $\overline{\mathscr{L}}$ and the position of its maximum represents the best estimate of the most likely values of the model parameters. Specifically, if the value of $q_{ir}$ is known, we can find the values of the model parameters $\pi$, $\theta$, $\phi$ where $\overline{\mathscr{L}}$ reaches its maximum. However, the calculation of $q_{ir}$ also requires the values of these model parameters. To address such a problem, an expectation-maximization (EM) algorithm is adopted.

Under the framework of EM algorithm, we first calculate the value of $q_{ir}$ by

$$q_{ir} = \frac{\Pr(A, g_i = r|\pi, \theta, \phi)}{\Pr(A|\pi, \theta, \phi)}$$

$$= \frac{\pi_r \prod_j \theta_{rj}^{A_{ij}^{+}} \phi_{rj}^{A_{ij}^{-}}}{\sum_s \pi_s \prod_j \theta_{sj}^{A_{ij}^{+}} \phi_{sj}^{A_{ij}^{-}}}. \tag{5.15}$$

Once we have the values of the $q_{ir}$, we can use them to evaluate the expected log-likelihood and hence to find the values of $\pi$, $\theta$, $\phi$ that maximize it.

Introducing the Lagrange multipliers $\rho$, $\gamma_r$ and $\lambda_r$ to incorporate the normalization conditions in Eq. 5.11, the expected log-likelihood expression to be maximized becomes

$$\widetilde{\mathscr{L}} = \overline{\mathscr{L}} + \rho\left(1 - \sum_r \pi_r\right) + \sum_r \gamma_r\left(1 - \sum_i \theta_{ri}\right) + \sum_r \lambda_r\left(1 - \sum_i \phi_{ri}\right). \tag{5.16}$$

By letting the derivative of $\widetilde{\mathscr{L}}$ to be 0, the maximum of the expected log-likelihood occurs at the places where

$$\pi_r = \frac{1}{n}\sum_i q_{ir},$$

$$\theta_{rj} = \frac{\sum_i A_{ij}^{+} q_{ir}}{\sum_{ik} A_{ik}^{+} q_{ir}}, \tag{5.17}$$

$$\phi_{rj} = \frac{\sum_i A_{ij}^{-} q_{ir}}{\sum_{ik} A_{ik}^{-} q_{ir}}.$$

When the algorithm converges, we obtain a set of values for hidden quantity $q_{ir}$ and model parameters $\pi$, $\theta$, $\phi$. This set of values is self-consistent with respect to Eqs. 5.15 and 5.17. However, it does not always correspond to the place where the log-likelihood reaches its maximum. In other words, the expectation-maximization algorithm may converge to local maxima of the log-likelihood. With different starting values, the algorithm may give rise to different solutions. To obtain a satisfactory solution, it is necessary to perform several runs with different starting values and then take the solution giving the highest log-likelihood over all the runs performed.

As indicated by Eqs. 5.15 and 5.17, a main property of our model is: without negative links, our model reduces to Newman's mixture model for the exploratory analysis of network structure. Actually, our model is also applicable to networks with either only positive links or only negative links.

Now we discuss the computational cost of the expectation-maximization algorithm for the fitting of our model. For each iteration in this algorithm, the cost consists in two parts. The first part is from the calculation of $q_{ir}$ using Eq. 5.15, whose time-complexity is $O(m \times c)$. Here $m$ is the number of edges in the network and $c$ is the number of groups. The second part is from the estimation of the model parameters using Eq. 5.17, whose time-complexity is also $O(m \times c)$. We use $T$ to denote the number of iterations before the iteration process converges. Then, the total cost of the expectation-maximization algorithm for our model is $O(T \times m \times c)$. It is difficult to give a theoretical estimation to the number $T$ of iterations. Generally speaking, $T$ is determined by the network structure and the starting values for the expectation-maximization algorithm.

In addition, our model assumes that the number $c$ of groups is known *a prior*. However, for many cases, this information is usually unknown. Thus it is desirable that there is a criterion to determine the appropriate group number for a given network. This task is known as the model selection issue in statistics. Several methods have been given to deal with this issue. We adopt the minimum description length principle, which is also used to handle the model selection issue in several existing generative models for network structure [11]. According to minimum description length principle, the required length to describe the network data is composed of two parts. The first part describes the coding length of the network using our model. This coding length is $-L$ for directed network and $-L/2$ for undirected network. The second part gives the length for coding model parameters. This part is $-\sum_{r,\pi_r>0} \ln \pi_r - \sum_{ri,\theta_{ri}>0} \ln \theta_{ri} - \sum_{ri,\phi_{ri}>0} \ln \phi_{ri}$. In this way, the optimal $c$ is the one which minimizes the total description length.

### *5.4.1 Comparison with Other Models*

In this section, we illustrate the connection or difference between our model and several existing models.

Firstly, we compare our model with Newman's mixture model, which is the basis of our model. Both our model and Newman's model define node groups according to similar connecting preference of nodes in the same group. This flexible definition enables these two models to explore multiple types of structural regularities in networks, including community structure and multi-partite structure. The main difference between these two models lies in whether they can deal with the coexistence of positive links and negative links. For Newman's mixture model, the negative links will cancel the positive links and will result in negative values for the parameters $\theta_{ri}$ and $q_{ir}$. This destroys the probabilistic interpretation of these two parameters. The possible reason for this problem is the hypothesis that there is only one unique probability $\theta_{ri}$ to control the connecting preference of nodes, which involves both positive and negative links. Such a problem facing Newman's mixture model disappears in our model by using two different probabilities, i.e., $\theta_{ri}$ and $\phi_{rj}$, to guide the formation of links, one for positive links and the other for negative links.

Secondly, we compare our model with existing methods for the detection of community structure in networks with positive and negative links. Two such kind of existing methods respectively generalize the modularity optimization method and the Potts model from networks with only positive links to networks with both positive and negative links [25, 26]. These methods can only detect community structure and are incapable of exploring other types of structural regularities in networks. This drawback is addressed in our model by the flexible definition of node group as a set of nodes with similar connecting patterns (similar friends or similar enemies) rather than as a set of highly connected nodes.

## *5.4.2 Experimental Results*

### 5.4.2.1 Test on Networks with Only Positive Links

Before testing our method on networks with negative and positive links, we first apply it to networks with only positive networks. Such a test aims to illustrate the capability of our method at exploring multiple types of structural regularities.

The first tested network is known to have assortative community structure and the other has disassortative multipartite structure. The first network is the social network of 62 bottlenose dolphins living in Doubtful Sound, New Zealand. This network is compiled by Lusseau [29] from seven years of field studies of the dolphins, with links between dolphins representing statistically significant frequent association. This network splits naturally into two groups during Lusseau's study.

Figure 5.6 shows the best division of the dolphin network into two groups found by our method with the group number $c = 2$. As we can see, all but one node is correctly assigned to their groups. In addition, our method also returns the preferences $\theta_{ri}$ for connections from nodes in group $r$ to each other node $i$. In Fig. 5.6, we represent by the size of nodes the probabilities $\theta_{1i}$ that links from nodes in group 1 will connect to node $i$. As we can see, the nodes with higher sizes are central to the group. Thus, $\theta_{ri}$ provides a measure for the importance of node $i$ for group $r$.

The second network is a bipartite network which is constructed from the Southern women dataset [30], depicting 18 women's attendance to 14 social events. In

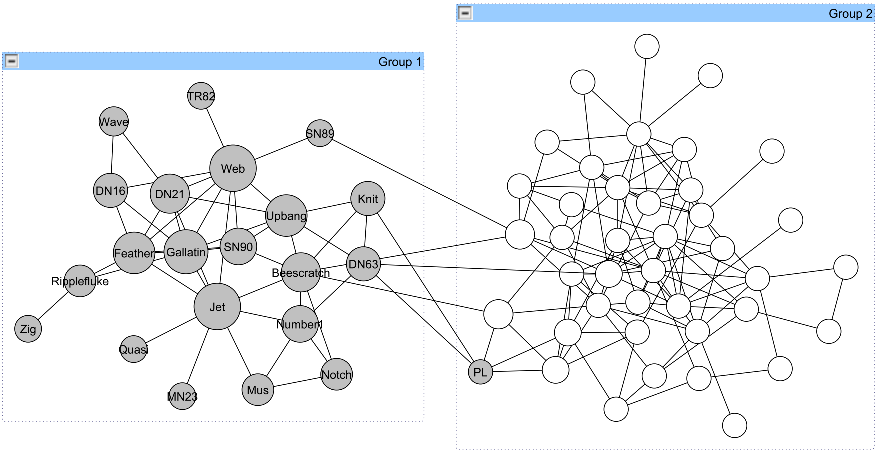


**Fig. 5.6** Results on the dolphin network. The real split of this network is represented by filled (or labeled) and unfilled (or unlabeled) circles. Two *shaded regions* are the groups identified by our method. The size of nodes represent the degree of preference being connected by nodes in group 1

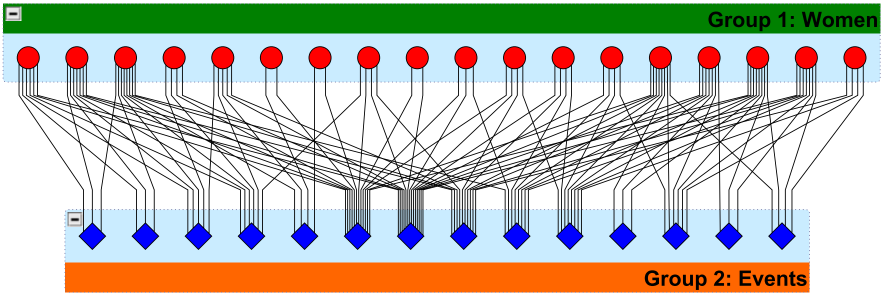
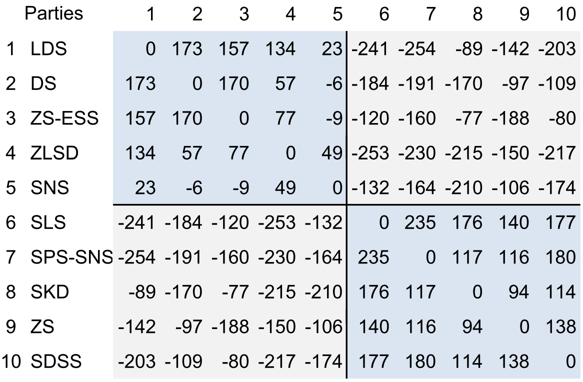


**Fig. 5.7** Result on the Southern women network. Here, *circles* represent women and *diamonds* represent social events attended by these women. Two *shaded regions* are the groups identified by our method

Fig. 5.7, we show the results of the application of our method to the women-event network. The bipartite structure of this network is exactly uncovered by our method. The key point to notice is that our method detects the bipartite structure without being told that it is to look for bipartite structure.

The above tests demonstrate that our method can find both the assortative community structure and the disassortative multipartite structure. This result is attributed to that our model is a generalization of Newman's mixture model, which possesses an important strength of our method, i.e., it is able to detect broad types of structural regularities without knowing in advance what type of structure to find. This advantage comes from the flexible definition of node group which relies on the similar connecting preference of group members to other nodes rather than on high link density within group. This basic idea is consistent with the social balance theory.



| | Parties | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | LDS | 0 | 173 | 157 | 134 | 23 | -241 | -254 | -89 | -142 | -203 |
| 2 | DS | 173 | 0 | 170 | 57 | -6 | -184 | -191 | -170 | -97 | -109 |
| 3 | ZS-ESS | 157 | 170 | 0 | 77 | -9 | -120 | -160 | -77 | -188 | -80 |
| 4 | ZLSD | 134 | 57 | 77 | 0 | 49 | -253 | -230 | -215 | -150 | -217 |
| 5 | SNS | 23 | -6 | -9 | 49 | 0 | -132 | -164 | -210 | -106 | -174 |
| 6 | SLS | -241 | -184 | -120 | -253 | -132 | 0 | 235 | 176 | 140 | 177 |
| 7 | SPS-SNS | -254 | -191 | -160 | -230 | -164 | 235 | 0 | 117 | 116 | 180 |
| 8 | SKD | -89 | -170 | -77 | -215 | -210 | 176 | 117 | 0 | 94 | 114 |
| 9 | ZS | -142 | -97 | -188 | -150 | -106 | 140 | 116 | 94 | 0 | 138 |
| 10 | SDSS | -203 | -109 | -80 | -217 | -174 | 177 | 180 | 114 | 138 | 0 |

**Fig. 5.8** The adjacency matrix of the parties of Slovene Parliamentary. This matrix characterize the relations among the 10 parties. The two *crossed lines* separate the two groups identified by our method

#### 5.4.2.2 Tests on Networks with Positive and Negative Links

We further test our method by applying it to several small real networks containing both positive and negative links. The community structure of these networks is also known. The first network is the network of 10 parties of the Slovene Parliamentary in 1994 [31]. This network is weighted and the weights characterize the distance between different parties. The weights were estimated by the 72 members of all the 90 members of the Slovene National Parliament by completing the questionnaire designed by a group of experts on Parliament activities. In the questionnaire, the respondents are required to estimate the distance between all the 45 pairs of parties on the scale from $-3$ to 3. These values respectively represent the pair of parties being "very dissimilar", "quite dissimilar", "dissimilar", "neither dissimilar nor similar (somewhere in between)", "similar", "quite similar", and "very similar". The final weights are the averaged value and multiplied by 100. Figure 5.8 depicts the adjacency matrix of the obtained network. Applying our method to this signed network, we show the results in Fig. 5.8, which is consistent with the real division among the parties in Slovene Parliamentary.

The second network is the Gahuku-Gama Subtribes network, which was created based on Read's study on the cultures of Eastern Central Highlands of New Guinea [32]. This network describes the political alliance and enmities among the 16 Gahuku-Gama subtribes, which are distributed in a particular area and are engaged in warfare with one another in 1954. The positive and negative links of the network correspond to political arrangements with positive and negative ties, respectively. Using our method, we analyze the structure of this signed network and detect three communities. This result is consistent with Read's study on this network.

#### 5.4.2.3 Application to the Network on International Relations

To illustrate the benefit brought by leveraging the negative links, we further compare our method to the existing community detection methods for signed networks and

Newman's method for exploratory analysis of structural regularities for networks with only positive links.

The used network is the network of international relations taken from the Correlates of War data set over the period 1993–2001 [26]. In this network, positive links represent military alliances and negative links denote military disputes. The disputes are associated with three hostility levels, from "no militarized action" to "interstate war". For each pair of countries, we chose the mean level of hostility between them over the given time interval as the weight of their negative link. The positive links denote the alliances: 1 for entente, 2 for non-aggression pact and 3 for defence pact. Finally, we normalized both the negative links and positive links into the interval [0, 1] and the final weight of the link among each pair of countries is the remainder of the weight of the normalized positive links subtracting the weight of the normalized negative links. The obtained network contains a giant component consisting of 161 nodes (countries) and 2517 links (conflicts or alliances). Here, we only investigate the structure of the giant component.

The community structure of this network has been investigated in several existing literatures. These existing studies indicated that there are six main *power blocs*, each power bloc consisting of a set of countries with similar actions of alliances or disputes. In [26], the authors labeled these power blocs as (1) The West; (2) Latin America; (3) Muslim World; (4) Asia; (5) West Africa; and (6) Central Africa. By setting the group number to be 6 and applying our method to this network, as shown in Fig. 5.9, we obtain similar results to the ones obtained in [26]. However, one notable difference exists between the two results. Specifically, the United States of America and Canada, the two counties in North America, are identified as a power bloc by our method while they are assigned to the West power bloc by the method in [26]. Such a difference is attributed to the very different assumptions behind the two methods. For the method in [26], the positive links are desired to connect the nodes within the same group while the negative links are among the nodes from different groups. For our method, the nodes in the same group have similar connecting preference to other nodes, i.e., they have common friends or common enemies. For the studied network, as shown in Fig. 5.9, the United States of America and Canada both have many positive links to the countries in the West and the Latin America while there are few positive links connecting countries belonging to the West and the Latin America. Such kind of connecting patterns can be correctly identified by our method. This further indicates that our method possesses distinct advantages over the existing methods designed for detecting assortative groups.

Furthermore, to illustrate the benefits provided by leveraging the negative links, we compare our method with the Newman's exploratory method which only takes into account the positive links. Our method and Newman's method give rise to similar results when applied on the international relationship network. However, two kinds of differences between the results of these two methods provide clear evidence to the advantage of our method. The first kind of difference is related to the nodes with only negative links or very few positive links. Our method can assign these nodes to appropriate groups by leveraging the information provided by negative links. Newman's method misclassifies these nodes because of its incapability to

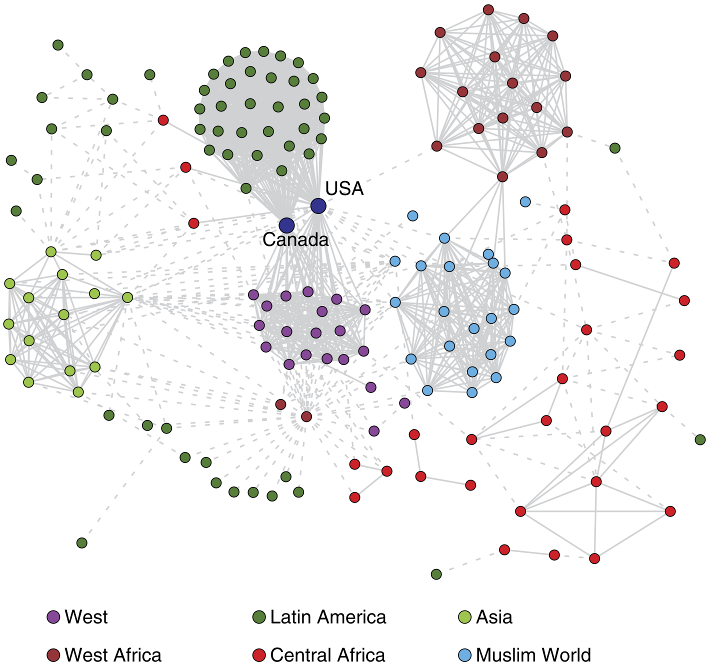


**Fig. 5.9** Results on the network of international relations taken from the Correlates of War. Alliances are represented by *solid lines* and disputes are represented by *dashed lines*. Colors differentiate the group of countries identified by our method

deal with negative links. The second kind of difference lies in the benefit provided by the negative links connecting different groups. For Newman's method, positive links connecting different groups will blur the boundary of the groups. For our method, the negative links connecting different groups can help distinguish different groups.

Limited by the lack of large-scale networks with both positive and negative links, we only illustrate the promising benefits possessed by our method through testing our method on several small real world networks with known structural regularities. We will conduct more extensive tests if we have large-scale networks in the future.

## 5.5 Conclusions

In this chapter, we consider the problem of exploring structural regularities of networks by dividing the nodes of a network into groups such that the members of each group have similar patterns of connections to other groups. Specifically, we propose a general statistical model to describe network structure. In this model, group is viewed as hidden or unobserved quantity and it is learned by fitting the observed network data using the expectation-maximization algorithm. Compared with existing models, the most prominent strength of our model is the high flexibility. This strength enables it to possess the advantages of existing models and to overcome their shortcomings in a unified way. As a result, not only broad types of structure can be detected without prior knowledge of the type of intrinsic regularities existing in the target network, but also the type of identified structure can be directly learned from the network. Moreover, by differentiating outgoing edges from incoming edges, our model can detect several types of structural regularities beyond competing models. Tests on a number of real world and artificial networks demonstrate that our model outperforms the state-of-the-art model at shedding light on the structural regularities of networks, including the overlapping community structure, multipartite structure and several other types of structure which are beyond the capability of existing models.

Furthermore, by generalizing Newman's mixture model according to social balance theory, we have studied the exploration of intrinsic structural regularities in networks with both positive and negative links. Without prior knowledge about which type of structure we are looking for, our method can detect both the assortative and disassortative structural regularities in a unified way. This is a distinct advantage of our method over existing methods designed either for certain type of structure or for networks with only positive links. Tests on a number of real world networks demonstrate the effectiveness and flexibility of our method. We look forward to seeing the applications of our method to more real world networks from various fields.

## References

1. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**, 7821–7826 (2002)
2. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
3. Flake, G.W., Lawrence, S.R., Giles, C.L., Coetzee, F.M.: Self-organization and identification of Web communities. IEEE Comput. **35**, 66–71 (2002)
4. Cheng, X.Q., Ren, F.X., Zhou, S., Hu, M.B.: Triangular clustering in document networks. New J. Phys. **11**, 033019 (2009)
5. Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Nature **433**, 895–900 (2005)
6. Cheng, X.Q., Ren, F.X., Shen, H.W., Zhang, Z.K., Zhou, T.: Bridgeness: A local index on edge significance in maintaining global connectivity. J. Stat. Mech. P10011 (2010)
7. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
8. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**, 036104 (2006)
9. Ramasco, J.J., Mungan, M.: Inversion method for content-based networks. Phys. Rev. E **77**, 036122 (2008)
10. Vazquez, A.: Population stratification using a statistical model on hypergraphs. Phys. Rev. E **77**, 066106 (2008)
11. Ren, W., Yan, G.Y., Liao, X.P., Xiao, L.: Simple probabilistic algorithm for detecting community structure. Phys. Rev. E **79**, 036111 (2009)
12. Zhang, H.Z., Qiu, B.J., Giles, C.L., Foley, H.C., Yen, J.: An lda-based community structure discovery approach for large-scale social networks. In: Proceedings of the IEEE Conference on Intelligence and Security Informatics, pp. 200–207 (2007)

13. Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature **453**, 98–101 (2008)
14. Newman, M.E.J., Leicht, E.A.: Mixture models and exploratory analysis in networks. Proc. Natl. Acad. Sci. USA **104**, 9564–9569 (2007)
15. Karrer, B., Newman, M.E.J.: Stochastic blockmodels and community structure in networks. Phys. Rev. E **83**, 016107 (2011)
16. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. J. Mach. Learn. Res. **9**, 1981–2014 (2008)
17. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E **64**, 026118 (2001)
18. Ball, B., Karrer, B., Newman, M.E.J.: Efficient and principled method for detecting communities in networks. Phys. Rev. E **84**, 036103 (2011)
19. Shen, H.W., Cheng, X.Q., Guo, J.F.: Exploring the structural regularities in networks. Phys. Rev. E **84**, 056111 (2011)
20. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**, 452–473 (1977)
21. Rosvall, M., Bergstrom, C.T.: An information-theoretic framework for resolving community structure in complex networks. Proc. Natl. Acad. Sci. USA **104**, 7327–7331 (2007)
22. Szell, M., Lambiotte, R., Thurner, S.: Multirelational organization of large-scale social networks in an online world. Proc. Natl. Acad. Sci. USA **107**, 13636–13641 (2010)
23. Heider, F.: Attitudes and cognitive organization. J. Psychol. **21**, 107–112 (1946)
24. Marvel, S.A., Kleinberg, J., Kleinberg, R.D., Strogatz, S.H.: Continuous-time model of structural balance. Proc. Natl. Acad. Sci. USA **108**, 1771–1776 (2011)
25. Gómez, S., Jensen, P., Arenas, A.: Analysis of community structure in networks of correlated data. Phys. Rev. E **80**, 016114 (2009)
26. Traag, V.A., Bruggeman, J.: Community detection in networks with positive and negative links. Phys. Rev. E **80**, 036115 (2009)
27. Rubinov, M., Sporns, O.: Weight-conserving characterization of complex functional brain networks. NeuroImage **56**, 2068–2079 (2011)
28. Yang, B., Cheung, W.K., Liu, J.M.: Community mining from signed social networks. IEEE Trans. Knowl. Data Eng. **19**, 1333–1348 (2007)
29. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait? Behav. Ecol. Sociobiol. **54**, 396–405 (2003)
30. Guimerà, R., Sales-Pardo, M., Amaral, L.A.N.: Module identification in bipartite and directed networks. Phys. Rev. E **76**, 036102 (2007)
31. Kropivnik, S., Mrvar, A.: An analysis of the Slovene parliamentary parties network. In: Developments in Statistics and Methodology, pp. 209–216 (1996)
32. Doreian, P., Mrvar, A.: A partitioning approach to structural balance. Soc. Netw. **18**, 149–168 (1996)