Statistica Sinica Preprint No: SS-2015-0040R2	
Title	A Hypothesis Testing Framework for Modularity Based Network Community Detection
Manuscript ID	SS-2015-0040.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0040
Complete List of Authors	Jingfei Zhang and Yuguo Chen
Corresponding Author	Yuguo Chen
E-mail	yuguo@illinois.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

A HYPOTHESIS TESTING FRAMEWORK FOR MODULARITY BASED NETWORK COMMUNITY DETECTION

Jingfei Zhang and Yuguo Chen

University of Miami and University of Illinois at Urbana-Champaign

Abstract: A relevant feature of networks is community structure. Detecting communities is of great importance in understanding, analyzing, and organizing networks, as well as in making informed decisions. Many approaches have been proposed for detecting community structure in networks, but few methods have been proposed for testing the statistical significance of detected community structures. In this paper, we describe a statistical framework for modularity-based network community detection. Under this framework, a hypothesis testing procedure is developed to determine the significance of an identified community structure. The proposed modularity is shown to be consistent under a degree-corrected stochastic block model framework. Several synthetic and real networks are used to demonstrate the effectiveness of our method.

Key words and phrases: Community detection, consistency, degree-corrected stochastic block model, hypothesis testing, modularity function.

1. Introduction

Networks have been the focus of much recent attention since they describe a multitude of complex systems found in such fields as biology, social science, information technology, finance, and many others. Networks are built upon nodes and the edges (or interactions) between them. For example, social networks consist of individuals and such interactions between them as friendship, collaboration, or similar personal interest. The World Wide Web describes the web pages and

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

their linking patterns. A stock market network models the stocks and their synchronized price fluctuations over time.

Existing networks often display a high level of local inhomogeneity, with high edge density within certain groups of nodes and low edge density between these groups. This feature is often referred to as "community structure" (Fortunato (2010)). Communities occur in many network systems in social science, biology, political science, economics, computer science, and other areas. In the protein-protein network, communities are groups of proteins that carry specific functions in the cell (Chen and Yuan (2006)); in the World Wide Web, communities correspond to groups of pages that are related to the same or similar topics (Dourisboure et al. (2007)).

Studying community structures can help us better understand networks, since nodes in the same community usually share common properties. For example, the members in a Facebook friendship community usually share similar demographic attributes or personal interests (Yang et al. (2013)), while different communities can exhibit significantly different network properties, which makes studying them at the level of the entire network inappropriate (Newman (2006b)). Community detection has applications. Thus identifying communities of customers with similar interests in the purchase relationship network of an online retail store can help in setting up more efficient recommendation systems (Reddy et al. (2002)).

Due to the importance of finding community structures in networks, there has been work on this topic in such fields as computer science, physics, statistics, and sociology (Agrawal and Kempe (2008); Reichardt an Bornholdt (2006); Newman and Girvan (2004); Snijders and Nowicki (1997)). Detecting communities in a network is not a trivial task. The number of possible partitions of the network is usually very large, especially when the number and the sizes of the communities are in general unknown. In this paper, community detection refers to partitioning the network

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

into several non-overlapping subnetworks. The terms community detection and network partition will be used interchangeably. See Fortunato (2010) for a review of the techniques for detecting overlapping communities.

Community detection approaches can be loosely divided into two classes. One involves maximizing a quality function over all possible partitions of the network. This includes the well-studied cut models (Flake et al. (2000)), spectral clustering (Shi and Malik (2000)), and modularity maximization (Newman and Girvan (2004)). Another class of techniques are model-based, fitting probabilistic models to the networks with community structures. This class includes the stochastic block model (Nowicki and Snijders (2001); Bickel and Chen (2009)), the degree-corrected stochastic block model (Karrer and Newman (2011); Jin (2015)), the mixed membership model (Airoldi et al. (2008)), and the multivariate latent variable model (Handcock et al. (2007)). From an algorithmic perspective, many model-based approaches lead to maximizing certain criteria, such as maximizing the profile likelihood over all possible partitions (Bickel and Chen (2009); Zhao et al. (2012)).

Formally defining community in a network is difficult, similar to defining cluster in multivariate analysis (Zhao et al. (2012)). Several ways to define a community have been proposed. One requires that two nodes in the same community are *stochastically equivalent* in the sense that exchanging the labels of these two nodes does not affect the probability of any event pertaining to the network (Fienberg et al. (1985)). This definition is used in most of the model-based approaches, such as the stochastic block model, degree corrected stochastic block model, and mixed membership model. A community, as adopted in this paper, has more edges between the nodes within a community and fewer edges between a community and the rest of the network. This definition is widely used in such algorithm-based community detection methods as modularity maximization algorithms, spectral clustering, and minimum-cut approaches. See Fortunato (2010) for a comprehensive review

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

of other definitions of communities.

Among community detection approaches, modularity maximization is one of the most popular (Fortunato (2010)). In Newman and Girvan (2004), a quality function called modularity was proposed to measure the quality of a network partition. Subsequent work has shown empirically that partitions that maximize the modularity function often identify interesting community structures (Newman (2004, 2006a, 2006b); Clauset et al. (2004); Chen and Yuan (2006)). However, the Newman-Girvan modularity function can be misleading. It has been shown that some random graphs with no community structures have partitions with large modularity values (Guimera et al. (2004); Reichardt and Bornholdt (2006)). Since the null model in the Newman-Girvan modularity function lacks a solid statistical basis, it is difficult to determine the statistical significance of the community structure obtained from maximizing the modularity function.

In this paper, we re-examine the null model in the Newman-Girvan modularity function and provide a statistical framework for modularity-based community detection. Based on it, we introduce a hypothesis testing procedure to determine the significance of the partitions obtained from maximizing the modularity function. We show that the modularity formulated under our framework is consistent under a degree-corrected stochastic block model framework.

The rest of the paper is organized as follows. Section 2 discusses the Newman-Girvan modularity function and its connection to the community structure in networks. Section 3 introduces the statistical framework for modularity-based community detection. A hypothesis testing procedure is then proposed for testing the significance of an identified community structure. Section 4 gives the connection between the proposed statistical framework and the degree-corrected stochastic block model. We show the consistency of the modularity function under the degree-corrected stochastic block model. Section 5 uses synthetic and real networks to demonstrate the effectiveness of our

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

method. Section 6 provides some concluding remarks.

2. Modularity and Community Structure

A network (or graph) $G(V, E)$ with a set of n nodes V and a set of edges E can be represented by its adjacency matrix A, where $A_{ij} = 1$ if there is a link from node i to node j and 0 otherwise. The node degree d_i is the number of edges connected to node i. We are mainly concerned with simple graphs (undirected graphs with no self-loops or multiple edges). For simple graphs, the adjacency matrix A is a symmetric 0-1 matrix with a zero diagonal. The column sums of A are the same as the degree sequence $\boldsymbol{d} = (d_1, \ldots, d_n)$ of $G(V, E)$. Moreover, the total number of edges in G is $m = \sum_{i<j} A_{ij}$.

Newman and Girvan (2004) proposed a hierarchical algorithm, in which edges with the highest betweenness are removed recursively until the network breaks down from one community of n nodes into n communities of one node. This whole process can be represented by a dendrogram showing various possible partitions of the network.

To determine which partition is optimal, Newman and Girvan (2004) defined a quality measure Q referred to as the *modularity*. Given a graph $G(V, E)$ with n nodes and community assignment $\boldsymbol{e} = (e_1, \ldots, e_n)$, where $e_i \in \{1, \ldots, K\}$ is the community that node i belongs to, the Newman-Girvan modularity Q_{NG} is defined as

$$Q_{NG}(\boldsymbol{e}, G) = \frac{1}{2m} \sum_{i,j} [A_{ij} - E(A_{ij})]\, \delta(e_i, e_j), \tag{2.1}$$

where $\delta(r, s) = 1$ if $r = s$ and 0 otherwise. Here $E(A_{ij})$ is the expected number of edges between node i and node j under some null model with no community structure. The modularity function measures the "discrepancy" between the observed number of edges and the expected number of edges within the communities under the null model. If the number of edges in the communities is close to the expected value, Q_{NG} is close to 0. When Q_{NG} approaches 1, it indicates strong

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

community structure. In Newman and Girvan (2004), the partition in the dendrogram that has the largest Q_{NG} value is outputted as the community structure.

Such a measure Q_{NG} of network partitions leads to a new class of approaches in community detection. Rather than just using Q_{NG} as a measure of the quality of a partition, one can instead directly try to find the partition that maximizes it. Brandes et al. (2008) showed that finding the partition that maximizes the modularity function for a given graph is NP-complete. Numerous heuristic approaches have been proposed to find partitions that maximize the Newman-Girvan modularity function (Newman (2006a); Agrawal and Kempe (2008); Reichardt and Bornholdt (2006); Clauset et al. (2004); Massen and Doye (2005); Wang et al. (2008)). This is still an active research topic.

After the partition that maximizes Q_{NG} is obtained, the interpretation of the result is important. Newman and Girvan (2004) suggested that networks with strong community structure typically have maximum modularity value $\max_{\boldsymbol{e}} Q_{NG}(\boldsymbol{e}, G) \in [0.3, 0.7]$. This is widely used as a rule of thumb in subsequent work. However, in general, a large Newman-Girvan modularity value does not necessarily indicate a community structure. Random graphs from the Erdös-Rényi model can have partitions with large Newman-Girvan modularity values (Guimera et al. (2004); Reichardt and Bornholdt (2006)), although such graphs are not supposed to have community structures, because the probability of having an edge between any pair of nodes is the same and every node is treated equally.

Based on the definition of modularity, a network should only be considered to have community structure if its maximized modularity value is significantly larger than the maximized modularity value of graphs from the null model. Testing the significance of a community structure in a network requires a well-defined null model. The null model in the Newman-Girvan modularity is, however,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

not clearly defined. Newman (2006b) discussed the importance of preserving the observed degree sequence in the null model and proposed setting the expected node degree $E(d_i)$ equal to the observed node degree d_i in the null model,

$$\sum_j E(A_{ij}) = d_i. \tag{2.2}$$

He also proposed that the edges be placed entirely at random in the null model. The probability that two nodes are placed at the ends of an edge should only depend on the degrees of the nodes,

$$E(A_{ij}) = f(d_i)f(d_j), \tag{2.3}$$

for some function $f(\cdot)$ (Newman (2006b)). It is easy to show that constraints (2.2) and (2.3) imply

$$E(A_{ij}) = \frac{d_i d_j}{2m}. \tag{2.4}$$

Thus the expectation $E(A_{ij})$ in the Newman-Girvan modularity is calculated without clearly specifying the null model. Here the expected number of edges $E(A_{ij})$ in (2.4) can be larger than one and the expected number of self links $E(A_{ii})$ can be larger than zero, as multiple edges and self-loops are allowed in this formulation.

As a null model in the Newman-Girvan modularity is not clearly specified, we describe a statistical framework, that includes a well-defined null model for modularity-based community detection.

3. Significance Testing in Modularity Based Community Detection

Given a graph $G(V, E)$ with n nodes and degree sequence $\boldsymbol{d} = (d_1, \ldots, d_n)$, the null model for the modularity measure is a random graph model with no community structure, and the null model should specify a distribution over the space of such.

The graphs in the null space should share some basic structural properties with G. Often, the distribution of the edges is highly inhomogeneous with global inhomogeneity, many vertices with

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

low degrees and a few vertices with high degrees, and local inhomogeneity, a high concentration of edges within certain groups of nodes and a low concentration of edges between these groups (Fortunato (2010)). To study local inhomogeneity, it is desirable to preserve the observed degree sequence in the null model. We fix the degree sequence of graphs from the null model at $\boldsymbol{d}$, and suppose they do not contain self-loops or multiple edges. In the following, our null space $\Sigma_{\boldsymbol{d}}$ is the set of all simple graphs with degree sequence $\boldsymbol{d} = (d_1, \ldots, d_n)$.

We assume that there is no preference for any graph in the null space $\Sigma_{\boldsymbol{d}}$ and take

$$p(g) = \frac{1}{|\Sigma_{\boldsymbol{d}}|}, \qquad g \in \Sigma_{\boldsymbol{d}}, \tag{3.1}$$

where $|\Sigma_{\boldsymbol{d}}|$ is the total number of graphs in $\Sigma_{\boldsymbol{d}}$. Section 4 discusses another motivation for the null model.

The proposed null model has some connection to, and several advantages over, the *configuration model* (Bender and Canfield (1978); Bollobás (1980)). In the configuration model, one assigns half edges to the vertices according to the degree sequence $(d_1, \ldots, d_n)$, and then performs a random matching of the $2m$ half edges. The outcomes of the configuration model are guaranteed to have the prescribed degree sequence $(d_1, \ldots, d_n)$. But, since the half-edges are matched randomly, the graphs produced may contain self-loops and multiple edges. Although removing graphs with multiple edges and self-loops leads to the uniform distribution over simple graphs, it deviates from the configuration model and the probability of having multiple edges and self-loops increases rapidly when the degrees increase (Cafieri et al. (2010); Chung and Lu (2002)).

Under the null model (3.1),

$$\sum_{j=1}^{n} E(A_{ij}) = d_i, \qquad i = 1, \ldots, n, \tag{3.2}$$

which means (2.2) is satisfied in our null model. Based on the null distribution (3.1), we revise

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

Newman-Girvan modularity function as

$$Q(\boldsymbol{e}, G) = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - E_{p,\Sigma_{\boldsymbol{d}}}(A_{i,j}) \right] \delta(e_i, e_j), \tag{3.3}$$

where the expectation $E_{p,\Sigma_{\boldsymbol{d}}}(\cdot)$ is taken with respect to $p(\cdot)$ on $\Sigma_{\boldsymbol{d}}$ given in (3.1).

To calculate $P_{ij} = E_{p,\Sigma_{\boldsymbol{d}}}(A_{ij})$, we notice that

$$P_{ij} = \frac{|\Sigma_{\boldsymbol{d}|A_{ij}=1}|}{|\Sigma_{\boldsymbol{d}}|} = 1 - \frac{|\Sigma_{\boldsymbol{d}|A_{ij}=0}|}{|\Sigma_{\boldsymbol{d}}|}, \tag{3.4}$$

where $|\Sigma_{\boldsymbol{d}|A_{ij}=k}|$ is the total number of simple graphs with degree sequence $\boldsymbol{d}$ and $A_{ij} = k$, $k = 0, 1$. Calculating $|\Sigma_{\boldsymbol{d}|A_{ij}=k}|$ and $|\Sigma_{\boldsymbol{d}}|$ is a difficult problem. Bender and Canfield (1978) derived an asymptotic formula for $|\Sigma_{\boldsymbol{d}|A_{ij}=0}|$ and $|\Sigma_{\boldsymbol{d}}|$ as $m \to \infty$ under uniformly bounded $\max_i d_i$. Mckay (1985) improved the asymptotic formula for $|\Sigma_{\boldsymbol{d}|A_{ij}=0}|$ and $|\Sigma_{\boldsymbol{d}}|$ and allow $\max_i d_i$ to increase with n. We use Mckay's (1985) asymptotic formula for $|\Sigma_{\boldsymbol{d}}|$ to derive an approximation to P_{ij}.

Theorem 1. *Let $G(V, E)$ be a simple graph with degree sequence $\boldsymbol{d} = (d_1, \ldots, d_n)$ with $m = \frac{1}{2}\sum_{i=1}^{n} d_i$ the total number of edges in $G(V, E)$. If $\max_i d_i = o(m^{1/4})$, as $n \to \infty$, P_{ij} is uniformly*

$$1 - e^{-\frac{d_i d_j}{2m} + o(1)}. \tag{3.5}$$

We refer to the supplementary material for the proof. In the theorem, since $\max_i d_i \geq 2m/n$, the condition $\max_i d_i = o(m^{1/4})$ implies that $\max_i d_i = o(n^{1/3})$, which describes the maximum degree of the graph. Moreover, it also implies that $\max_{i,j} \frac{d_i d_j}{2m} \to 0$, which leads to

$$1 - e^{-\frac{d_i d_j}{2m}} = \frac{d_i d_j}{2m} + o\Big(\frac{d_i d_j}{2m}\Big).$$

Corollary 1. *Under the conditions of Theorem 1 we have, as $n \to \infty$, P_{ij} is uniformly*

$$\frac{d_i d_j}{2m} + o(1). \tag{3.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

Newman-Grivan modularity also uses $d_i d_j/2m$ to approximate $E(A_{ij})$ (see (2.4)). One major difference is that our approximation is based on a well-defined null hypothesis, while (2.4) in Newman-Grivan modularity is based on a set of constraints based on expected degrees.

Approximating the linking probability $P_{ij} = E_{p,\Sigma_{\boldsymbol{d}}}(A_{ij})$ based on (3.5) or (3.6) may not be satisfactory when the condition in Theorem 1 is not satisfied, or when the graph is of moderate size. In that case, Monte Carlo methods can be used to estimate P_{ij}. The null distribution (3.1) is the uniform distribution over graphs with fixed degree sequence. If we generate N random samples $g_1, \ldots, g_N$ from the null distribution specified in (3.1), then P_{ij} can be estimated by

$$P_{ij} = \sum_{l=1}^{N} A_{ij}^{(l)}/N, \tag{3.7}$$

where $A_{ij}^{(l)}$ is the (i,j)-th entry of the adjacency matrix of g_l.

One way to generate graphs with fixed degree sequence uses the configuration model, but the graphs generated may contain self-loops and multiple edges. Discarding the graphs with self-loops and multiple edges can waste a lot of samples since the probability of having multiple edges and loops increases quickly when the degrees increase (Cafieri et al. (2010); Chung and Lu (2002)).

A Markov chain Monte Carlo (MCMC) algorithm, often referred to as the "rewiring" method, provides an easy way to sample from the null distribution (Blitzstein and Diaconis (2010)). Other more sophisticated schemes for estimating P_{ij}, such as sequential importance sampling, can be found in Blitzstein and Diaconis (2010) and Zhang and Chen (2013).

With a well-defined null hypothesis (3.1), we are able to introduce a hypothesis testing procedure that can test the significance of an identified structure. As discussed earlier, a network should only be considered to have community structure if its maximized modularity value is significantly larger than the maximized modularity value of graphs from the null model. Given a graph $G(V,E)$ and a community assignment $\boldsymbol{e}^* = (e_1^*, \ldots, e_n^*)$, the statistical significance of the partition is the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

probability that $Q(\boldsymbol{e}^*, G)$ is smaller than the the maximized modularity value of graphs from the null model,

$$P[Q(\boldsymbol{e}^*, G) \leq \max_{\boldsymbol{e}} Q(\boldsymbol{e}, g)], \tag{3.8}$$

where g follows the null distribution in (3.1). Our test statistic, the maximized modularity, tends to be large under the alternative of community structure. Therefore when the null is rejected, it indicates that the data favors the alternative of community structure.

A straightforward way to estimate the p-value in (3.8) is to generate samples $g_1, \ldots, g_N$ from the uniform distribution over $\Sigma_{\boldsymbol{d}}$ using the MCMC algorithm, and then approximate (3.8) by

$$\frac{1}{N}\sum_{i=1}^{N} I\left(Q(\boldsymbol{e}^*, G) \leq \max_{\boldsymbol{e}} Q(\boldsymbol{e}, g_i)\right). \tag{3.9}$$

The sequential importance sampling algorithms proposed in Blitzstein and Diaconis (2010) and Zhang and Chen (2013) can be used to approximate the p-value as well.

4. Connection to Degree-Corrected Stochastic Block Model

In this section, we use the degree-corrected stochastic block model (DCSBM) to provide another motivation for the proposed null model (3.1). We also show that the proposed modularity is consistent under the DCSBM.

The stochastic block model (SBM) is a widely used models for networks with communities. Consider a graph $G(V, E)$ with n nodes and adjacency matrix A. Under the stochastic block model, there are K classes (or blocks) such that each node belongs to only one of the classes. Let $\boldsymbol{c} = (c_1, \ldots, c_n)$ denote the true community labels, where c_i is the community to which node i belongs. Under the SBM, each A_{ij} is an independent Bernoulli random variable with

$$E[A_{ij}|\boldsymbol{c}] = W_{c_i c_j}, \tag{4.1}$$

where W_{rs} is the probability that a node in block r is linked to a node in block s.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

When fitted to an observed network, the SBM can uncover the underlying block (or community) structure, but it has limitations in its application. For every block in the model, within the same block all nodes are considered to be equivalent. Thus the model does not allow the hub nodes (nodes that have significantly more links than the others) that are frequently observed in real networks. Fitting a SBM to networks that have highly inhomogeneous degree distributions can lead to inaccurate results (Karrer and Newman (2011)). To address this problem, Karrer and Newman (2011) proposed the DCSBM, which adds parameters to account for the degree correction in the SBM. They showed empirically that when there is heterogeneity in the degree distribution, the DCSBM fits better than the standard SBM.

A generalized DCSBM with K blocks can be written as (Peng and Carvalho (2013))

$$A_{ij} \sim \text{Bernoulli}(q^{-1}(\theta_i + \theta_j + Z_{c_i c_j})), \tag{4.2}$$

where $q(\cdot)$ is a link function, $Z_{c_i c_j}$ reflects the linking probability between block c_i and block c_j, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is a vector of node specific parameters. With a logit link, the likelihood of the DCSBM can be written as

$$P(G|\boldsymbol{\theta}, Z) = \prod_{i<j} [\text{logit}^{-1}(\theta_i + \theta_j + Z_{c_i c_j})]^{A_{ij}} [1 - \text{logit}^{-1}(\theta_i + \theta_j + Z_{c_i c_j})]^{1-A_{ij}}. \tag{4.3}$$

If we choose the DCSBM as the underlying model for the network, the null model is naturally taken to be the DCSBM with only one block. We then have

$$A_{ij} \sim \text{Bernoulli}(\text{logit}^{-1}(\theta_i + \theta_j + Z)) \tag{4.4}$$

under the null model, where the Z matrix degenerates to a single parameter since there is only one block. This is essentially the well-studied logistic linear model (or β-model) for network data (Holland and Leinhardt (1981); Chatterjee et al. (2011); Perry and Wolfe (2012); Park and Newman

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

(2004); Blitzstein and Diaconis (2010)). Under the null model, the likelihood in (4.3) can be simplified to

$$P(G|\boldsymbol{\theta}, Z) = \frac{e^{\sum_i Zd_i/2 + \sum_i \theta_i d_i}}{\prod_{i<j}(1 + e^{\theta_i + \theta_j + Z})}. \tag{4.5}$$

Here (4.5) has the nuisance parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and Z. In hypothesis testing, one way to remove the effect of the nuisance parameters is to condition on the sufficient statistics (Lehmann (1986)), which are the degree sequence $\boldsymbol{d} = (d_1, \ldots, d_n)$ in model (4.5). Notice that conditioning on $\boldsymbol{d}$, the null model (4.5) becomes the uniform distribution over $\Sigma_{\boldsymbol{d}}$, the same as the null model proposed in (3.1).

4.1. Consistency Under the Degree-Corrected Stochastic Block Model

According to Bickel and Chen (2009), a community detection criterion Q is strongly consistent if $\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{e}} Q(\boldsymbol{e}, G)$ satisfies

$$P(\hat{\boldsymbol{c}} = \boldsymbol{c}) \to 1 \qquad \text{as } n \to \infty, \tag{4.6}$$

where $\boldsymbol{c}$ is the true community labels for the nodes. Zhao et al. (2012) defined weak consistency for a community detection criterion Q as

$$\forall \epsilon > 0, \qquad P\left[\left(\frac{1}{n}\sum_{i=1}^{n} I(\hat{c}_i \neq c_i)\right) < \epsilon\right] \to 1 \qquad \text{as } n \to \infty. \tag{4.7}$$

When the community detection criteria are invariant under permutations of the community labels, there are identifiability issues with these definitions. Zhao et al. (2012) suggested replacing $\hat{\boldsymbol{c}} = \boldsymbol{c}$ with $\hat{\boldsymbol{c}}$ and $\boldsymbol{c}$ belonging to the same equivalent class under permutation. In both types of consistency, the number of communities K is assumed to be known. Therefore, all modularity consistency properties are shown for

$$\hat{\boldsymbol{c}} = \arg \max_{\substack{\boldsymbol{e}=(e_1,\ldots,e_n) \\ e_i \in \{1,\ldots,K\}}} Q(\boldsymbol{e}, G). \tag{4.8}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

Under the proposed framework, assuming the graph satisfies the sparsity condition in Corollary 1, we detect communities by finding the maximizer of the modularity function

$$\hat{\boldsymbol{c}} = \arg \max_{\substack{\boldsymbol{e}=(e_1,\ldots,e_n) \\ e_i \in \{1,\ldots,K\}}} \frac{1}{2m} \sum_{i,j} \Big(A_{ij} - \frac{d_i d_j}{2m} \Big) \delta(e_i, e_j). \tag{4.9}$$

In Section 4, we use the DCSBM with a logit link to provide another motivation for the uniform null model (3.1). In this section, we show that (4.9) is consistent under the framework of the DCSBM with a logit link.

Bickel and Chen (2009) showed that the Newman-Girvan modularity is strongly consistent under the standard SBM with some constraints on the parameters. This result was extended by Zhao et al. (2012) to the DCSBM with a log link. To consider the consistency of (4.9) under the DCSBM with a logit link, we first give a slightly different formulation of the DCSBM with a logit link, following Zhao et al. (2012).

1. Each node i is associated with a pair of latent variables (c_i, θ_i), where $c_i \in \{1, \ldots, K\}$ and θ_i takes values in $x_1 \leq \cdots \leq x_N$.

2. The marginal distribution of $\boldsymbol{c} = (c_1, \ldots, c_n)$ follows a multinomial distribution with parameters $(\pi_1, \ldots, \pi_K)$.

3. Given $\boldsymbol{c}$ and $\boldsymbol{\theta}$, the edges A_{ij} are independent Bernoulli random variables with

$$\text{logit}(E[A_{ij}|\boldsymbol{c}, \boldsymbol{\theta}]) = \theta_i + \theta_j + Z_{c_i c_j},$$

where Z is a symmetric $K \times K$ matrix.

The parameter Z_{ij} captures the within and between community linking probabilities. Self loops are allowed in the model for simplicity. This makes no difference in the results compared to the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

setting where diagonal items are forced to be zero, but allowing self loops makes the notation much simpler. If we set all θ_i to zero, the above DCSBM with a logit link is the standard SBM.

Define $\Pi_{K\times N}$ to be the joint distribution of (c_i, θ_i), $P(c_i = a, \theta_i = x_u) = \Pi_{au}$. Here we do not assume that $\boldsymbol{\theta}$ and $\boldsymbol{c}$ are independent, since it is possible that the community labels and the node specific random variables are not independent in the model.

To ensure sparsity, we need to have P_{ij} scale with n, or else the graph is going to become unrealistically dense as $n \to \infty$ (Zhao et al. (2012)). Following Bickel and Chen (2009), we reparameterize P_{ij} as $P_{ij}^{(n)} = \rho_n P_{ij}$, where $\rho_n \equiv P(\text{Edge}) \to 0$ and P_{ij} is fixed as $n \to \infty$. We define the expected degree $\lambda_n \equiv E(\text{Degree}) = n\rho_n$. This reparameterization allows us to separate $\rho_n \propto E(\text{Degree})$ from the inhomogeneity structure of the graph. See Bickel and Chen (2009) for a more detailed explanation of the reparameterization.

Theorem 2. *Under the degree-corrected stochastic block model with a logit link, if for all $a \neq b$, the parameters satisfy*

$$logit^{-1}(x_u+x_v+Z_{ab}) < \frac{1}{H_0}\left(\sum_{b'v'} logit^{-1}(x_u + x_{v'} + Z_{ab'})\Pi_{b'v'}\right)\left(\sum_{a'u'} logit^{-1}(x_{u'} + x_v + Z_{a'b})\Pi_{a'u'}\right), \tag{4.10}$$

$$logit^{-1}(x_u + x_v + Z_{aa}) > \frac{1}{H_0}\left(\sum_{b'v'} logit^{-1}(x_u + x_{v'} + Z_{ab'})\Pi_{b'v'}\right)^2, \tag{4.11}$$

where $H_0 = \sum_{abuv} logit^{-1}(x_u + x_v + Z_{ab})\Pi_{au}\Pi_{bv}$, then the modularity in (4.9) is strongly consistent when $\lambda_n/logn \to \infty$, and weakly consistent when $\lambda_n \to \infty$.

See the supplementary material for the proof. The constraints (4.10) and (4.11) on the parameters in Theorem 2 essentially require that links are more likely to be established within communities than between communities. In the simplest case with $K = 2$ and there is no degree correction ($\theta_i = 0$), conditions (4.10) and (4.11) can be simplified to

$$\text{logit}^{-1}(Z_{11})\text{logit}^{-1}(Z_{22}) > \left[\text{logit}^{-1}(Z_{12})\right]^2. \tag{4.12}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

From the model specification, we have $P(A_{ij} = 1|c_i = a, c_j = b) = \text{logit}^{-1}(Z_{ab})$, where c_i is the community to which node i belongs. Then (4.12) can be expressed as

$$P(A_{ij} = 1|c_i = 1, c_j = 1)P(A_{ij} = 1|c_i = 2, c_j = 2) > [P(A_{ij} = 1|c_i = 1, c_j = 2)]^2 . \tag{4.13}$$

A DCSBM with the following constraints on the parameters will satisfy condition (4.13):

$$P(A_{ij} = 1|c_i = 1, c_j = 1) > P(A_{ij} = 1|c_i = 1, c_j = 2), \tag{4.14}$$

$$P(A_{ij} = 1|c_i = 2, c_j = 2) > P(A_{ij} = 1|c_i = 1, c_j = 2). \tag{4.15}$$

Constraints (4.14) and (4.15) correspond to the motivation that links are more likely to be established within communities than between communities.

The consistency result suggests that if the graphs are from a DCSBM with K communities, then the community labels obtained from maximizing the modularity function Q will be close to the true community labels as the number of nodes goes to infinity. By setting $\boldsymbol{\theta}$ to zero in Theorem 2, we obtain similar results for the SBM.

COROLLARY 2. *Under the stochastic block model with a logit link, if the parameters satisfy*

$$logit^{-1}(Z_{ab}) < \frac{1}{H_0}\left(\sum_{b'} logit^{-1}(Z_{ab'})\pi_{b'}\right)\left(\sum_{a'} logit^{-1}(Z_{a'b})\pi_{a'}\right), \tag{4.16}$$

$$logit^{-1}(Z_{aa}) > \frac{1}{H_0}\left(\sum_{b'} logit^{-1}(Z_{ab'})\pi_{b'}\right)^2, \tag{4.17}$$

where $H_0 = \sum_{ab} logit^{-1}(Z_{ab})\pi_a\pi_b$, *then the modularity in (4.9) is strongly consistent when* $\lambda_n/logn \to \infty$, *and weakly consistent when* $\lambda_n \to \infty$.

5. Numerical Examples

In this section, we denote the Newman-Girvan modularity function (2.1) by Q_{NG}. Our modularity function (3.3) calculated using the approximation (3.5) is referred to as the asymptotically

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

approximated modularity, denoted by Q_{asym}, and the modularity (3.3) calculated using the approximation (3.7) based on MCMC algorithms is referred to as the MCMC approximated modularity, denoted by Q_{MCMC}.

We consider two modularity maximization approaches. For small graphs with no more than a hundred nodes, we use the linear programming approach proposed in Agrawal and Kempe (2008). For graphs of moderate size, the algorithm runs fairly fast and has more accurate results compared to the approaches designed for large graphs. For large graphs, we propose a fast spectral clustering algorithm, a simplification of the algorithm discussed in Newman (2006b). We refer to the online supplementary material for the proposed spectral clustering method.

Our approach is not tied to the proposed spectral clustering method, and most existing modularity maximization methods can be used under our framework without modification. For example, the well-known Louvain method (Blondel et al. (2008)) can be used when the network is large. See Fortunato (2010) for a comprehensive review on modularity maximization methods and software/code sources. To sample from the null hypothesis, one can use the "rewire" function from the R package "igraph" (*http://igraph.org/r/*). In our simulation studies, 1,000,000 rewiring steps on a network with 1,000,000 nodes and 3,000,000 edges took about 8 seconds on an iMac desktop with 3.2 GHz quad-core processor.

The linear programming algorithm in Agrawal and Kempe (2008) is coded in C++ and implemented in a CPLEX environment. The fast modularity maximization algorithm in the online supplementary material is coded and implemented in R. All examples were run on a MacBook Pro with 2.26 GHz Intel Core 2 Duo processor.

5.1. Erdös-Rényi Random Graphs

For many networks, a high modularity value indicates a strong community structure, but this

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

is not true in general. It has been shown that random graphs can have partitions with large modularity values (Guimera et al. (2004); Reichardt and Bornholdt (2006)). To interpret the results from modularity-based community detection, it is necessary to test the significance of the maximized modularity value. In this section, we look at the maximized modularity function of Erdös-Rényi random graphs and demonstrate the use of our hypothesis testing procedure.

In an Erdös-Rényi graph, edges are established independently between each pair of nodes with equal probability p, thus they have no community structures. Graphs generated from the Erdös-Rényi model can have large modularity values. Figure 6.1 is the histogram of $\max_{\boldsymbol{e}} Q_{NG}(\boldsymbol{e}, G)$ of 100 Erdös-Rényi graphs with $n = 100$ and $p = 0.05$. The average, minimum, and maximum of the 100 maximum modularity $\max_{\boldsymbol{e}} Q_{NG}(\boldsymbol{e}, G)$ are 0.399, 0.350, and 0.451, respectively. Based on the general rule of thumb of Newman and Girvan (2004), all 100 Erdös-Rényi graphs are considered to have strong community structure since they have $\max_{\boldsymbol{e}} Q_{NG}(\boldsymbol{e}, G) \in [0.3, 0.7]$.

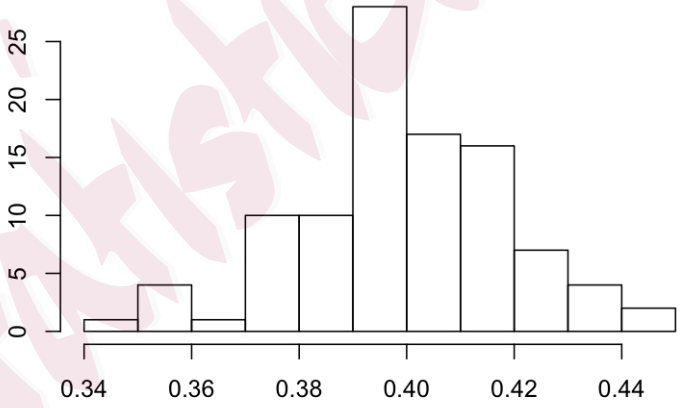

Figure 6.1: Histogram of $\max_{\boldsymbol{e}} Q_{NG}(\boldsymbol{e})$ for 100 Erdös-Rényi graphs.

To better understand the community detection results, it is useful to perform our test of hypothesis. Consider randomly generated Erdös-Rényi graphs $G_1, \ldots, G_{100}$ with $n = 100$ and $p = 0.05$. For each graph G_i, we calculated its maximized modularity $\max_{\boldsymbol{e}} Q_{asym}(\boldsymbol{e}, G_i)$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

$\max_{\boldsymbol{e}} Q_{MCMC}(\boldsymbol{e}, G_i)$. To perform hypothesis testing on each graph G_i, we generated 1000 samples $g_1^{(i)}, \ldots, g_{1000}^{(i)}$ uniformly from the set of simple graphs with the same degree sequence as G_i using MCMC. Then we estimated the p-value using

$$p_1^{(i)} = \frac{1}{1000} \sum_{j=1}^{1000} I\left(\max_{\boldsymbol{e}} Q_{asym}(\boldsymbol{e}, G_i) \leq \max_{\boldsymbol{e}} Q_{asym}(\boldsymbol{e}, g_j^{(i)})\right), \tag{6.1}$$

$$p_2^{(i)} = \frac{1}{1000} \sum_{j=1}^{1000} I\left(\max_{\boldsymbol{e}} Q_{MCMC}(\boldsymbol{e}, G_i) \leq \max_{\boldsymbol{e}} Q_{MCMC}(\boldsymbol{e}, g_j^{(i)})\right), \tag{6.2}$$

for $\max_{\boldsymbol{e}} Q_{asym}(\boldsymbol{e}, G_i)$ and $\max_{\boldsymbol{e}} Q_{MCMC}(\boldsymbol{e}, G_i)$.

Figure 6.2 shows the histograms for the 100 p-values for the two cases. Under the null model, the p-values are roughly uniformly distributed. If we set the significance level at 0.05, Type I error is estimated to be 0.03 and 0.07, respectively, for the two cases. This example shows that the test of hypothesis is needed in order to decide the significance of an identified community structure.

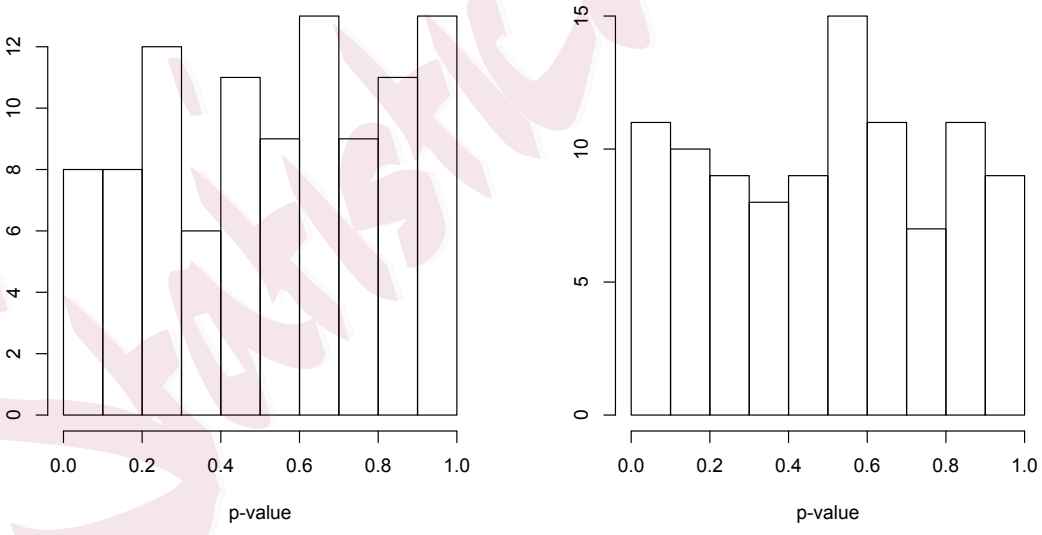

Figure 6.2: Histograms of 100 p_1-values (left) and 100 p_2-values (right).

We also ran the above simulation for Erdös-Rényi graphs with different choices of the number of nodes n and connection probability p. The Type I error for our test is estimated based

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

on $\max_{\boldsymbol{e}} Q_{asym}$ with significance level set at 0.05. The results summarized in Table 6.1 further demonstrate that the level of our proposed hypothesis testing procedure is well controlled.

$n = 100$		$n = 500$		$n = 1000$	
p	Type I error	p	Type I error	p	Type I error
0.05	7%	0.05	4%	0.05	2%
0.1	6%	0.1	4%	0.1	7%
0.25	3%	0.25	4%	0.25	5%
0.5	7%	0.5	5%	0.5	6%

Table 6.1: Type I error for Erdös-Rényi model with n nodes and edge probability p.

5.2. Synthetic Modular Networks

We tested our community detection procedure on networks known to have community structures. We generated 100 graphs from the standard stochastic block model with $n = 1000$ nodes and three blocks with sizes 200, 300, and 500. The linking probability matrix for the stochastic block model was set to $B = \begin{pmatrix} 0.5 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.1 \\ 0.1 & 0.1 & 0.2 \end{pmatrix}$. Here links are more likely to be established within the blocks and less likely to be established between the blocks, and graphs generated from this model should have community structure.

Figure 6.3 is the histogram of $\max_{\boldsymbol{e}} Q_{NG}(\boldsymbol{e}, G)$ for the 100 graphs generated from the model. As the maximum of the 100 values of $\max_{\boldsymbol{e}} Q_{NG}(\boldsymbol{e}, G)$ is 0.263, the general rule of thumb of Newman and Girvan (2004) suggests that none of the 100 graphs have strong community structures. Again the general rule of thumb can lead to false conclusions.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

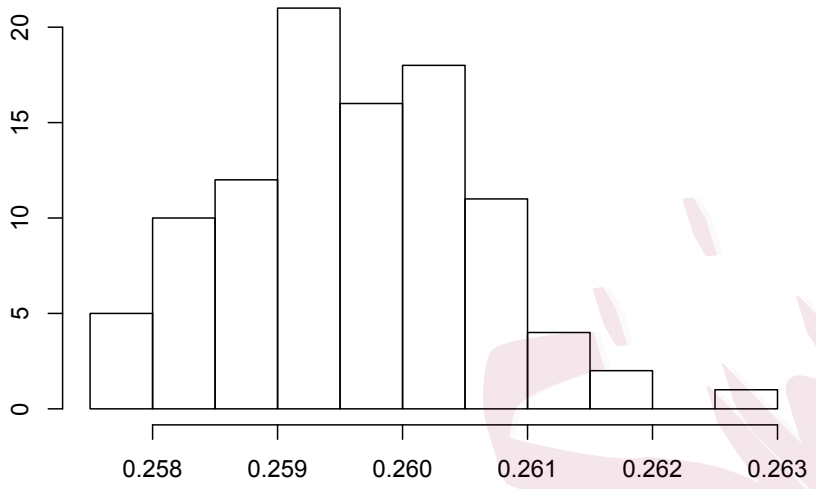

Figure 6.3: Histogram of the maximized modularity $\max_{\boldsymbol{e}} Q_{NG}(\boldsymbol{e})$ for 100 graphs generated from a stochastic block model with community structure.

Consider the same set of 100 graphs $G_1, \ldots, G_{100}$. Since the graph is large, we only computed Q_{asym}. For each graph G_i, we calculated $\max_{\boldsymbol{e}} Q_{asym}(\boldsymbol{e}, G_i)$. To perform hypothesis testing on each graph G_i, we generated 1000 samples $g_1^{(i)}, \ldots, g_{1000}^{(i)}$ from (3.1) using MCMC and estimated the p-value. Here we took a sample after every 20,000 (roughly a quarter of the total number of edges in the graph) MCMC steps to reduce the dependence between the samples.

For all 100 graphs, the p-values of the identified community structure were estimated to be 0, which suggests that all 100 graphs be considered to have strong community structures. This indicates that the proposed hypothesis testing procedure has good power in detecting communities. Moreover, for all 100 graphs, the community assignment has a misclassification rate of 0.

To further explore the power of our test, we did simulation studies on graphs generated from degree corrected stochastic blockmodels with 1000 nodes and two communities. Node labels were

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

generated independently with $P(c_i = 1) = \pi$ and $P(c_i = 2) = 1 - \pi$. By varying π, we can investigate the robustness of our method to unbalanced community sizes. The linking parameters matrix $Z = \begin{pmatrix} 1.5 & -1.5 \\ -1.5 & 1.5 \end{pmatrix}$ and the degree correction $\boldsymbol{\theta}$ was generated from a power law distribution with the exponent equal to 2.5. We generated edges with probability

$$P(A_{ij} = 1) = \rho \times \text{logit}^{-1}(Z_{c_i c_j} + \theta_i + \theta_j). \tag{6.3}$$

By varying ρ, we controlled the densities of the graph. Through $\pi = 0.5$, 0.25 and 0.1 and $\rho = 0.05$, 0.1, 0.25 and 0.5, we found no Type II errors (significance level set at 0.05).

5.3. Krebs' Network of Books on American Politics

The Krebs' network of books on American politics (available at *http://www.orgnet.com/*) has 105 vertices and 441 edges. Each node represents a book on US politics that is sold by the online bookseller Amazon.com. Each edge between a pair of nodes represents the frequent co-purchasing of the two books by the same buyers, which is indicated by the "customers who bought this book also bought these other books" feature on Amazon.com. Newman (2006a) provided a classification of these 105 books as liberal, conservative, or neutral, based on a reading of the descriptions and reviews of the books posted on Amazon.com.

This network is of moderate size and we only use the MCMC approximated modularity function Q_{MCMC}. Here Q_{MCMC} was calculated based on 1000 MCMC samples, each taken after 1000 MCMC moves. Using the modularity maximization algorithm of Agrawal and Kempe (2008), Q_{MCMC} is maximized at $K = 5$ with $\max Q_{MCMC} = 0.535$. When our test of hypothesis to decide the significance of the partition was run using the maximized Q_{MCMC} of the 1000 MCMC samples from the null model (3.1) of no community structure, the largest value seen was < 0.3, which indicates the identified community structure is significant.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

Among the five identified communities in Figure 6.4, two large communities are obviously the liberal community and the conservative community. Based on the members in the communities, the three smaller communities are roughly neutral, neutral conservative and conservative. One interesting observation is that the smaller conservative community on the rightmost is almost only connected to the large conservative community, and has almost no connections to the liberal community and neutral community. This indicates that if customers buy books from this small conservative community, it is very unlikely that they will buy books from the liberal community or the neutral community. Some examples of the books in this more extreme conservative community are "*Useful Idiots: How Liberals Got It Wrong in the Cold War and Still Blame America First*" by Mona Charen, "*The Right Man: The Surprise Presidency of George W. Bush*" by David Frum and "*The Savage Nation: Saving America from the Liberal Assault on Our Borders, Language and Culture*" by Michael Savage.

6. Discussion

In this paper, we provide a statistical framework for the modularity-based community detection. The proposed modularity function and statistical testing procedure perform well with both simulated and existing networks. We also show that under the degree-corrected stochastic block model, the proposed modularity function is consistent as a community detection criterion.

The modularity function Q can have negative values, the number of edges within communities is less than its expected value under the null model. A partition with large negative modularity suggests the existence of multipartite structure (Newman (2006b)). To detect the multipartite structure in the network, one can minimize the modularity function. The statistical framework proposed in the paper can be used to test the significance of an identified multipartite structure.

Besides modularity-based community detecting approaches, other methods have been proposed

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

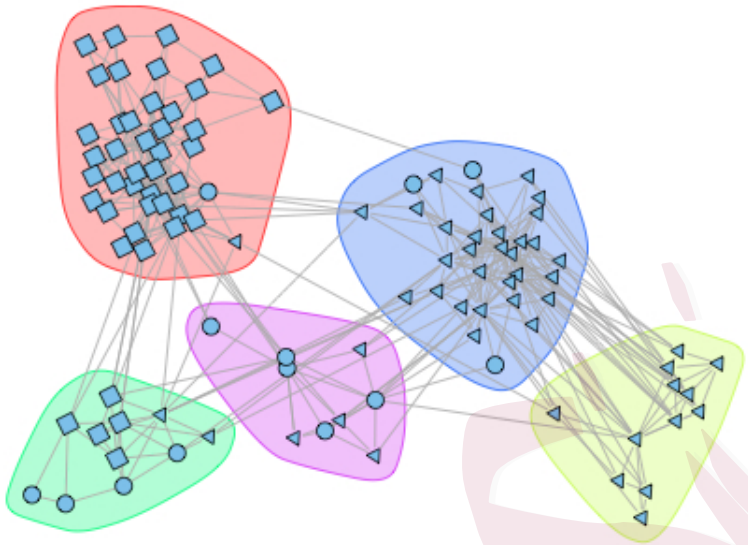

Figure 6.4: Krebs' network of books on American politics. Square nodes denote liberal books, triangle nodes denote conservative books and circle nodes denote neutral books.

for detecting communities in networks. However, not much work has been done to assess the statistical significance or quantify the uncertainty associated with identified community structures. Another work in our direction is Choi, Wolfe, and Airoldi (2011).

Extending our method to directed networks is an interesting research topic. A few methods have been proposed to look for communities in directed networks using modified modularity functions, see Fortunato (2012) for a review. To extend our method to directed networks, we need an appropriate null model and a modularity function for directed graphs, and new algorithms may be needed for finding optimal partitions.

Supplementary Material

The online supplementary material includes proofs of the theoretical results and details of the proposed modularity maximization algorithm.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

Acknowledgement

This work was partially supported by the National Science Foundation grants DMS-1106796 and DMS-1406455.

References

Agrawal, G. and Kempe, D. (2008). Modularity-maximizing graph communities via mathematical programming. *The European Physics Journal B* **66**, 409-418.

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008). Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research* **9**, 1981-2014.

Bender, E. and Canfield, R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory A* **24**, 296-307.

Bickel, P. and Chen, A. (2009). A non-parametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences* **106**, 21068-21073.

Blitzstein, J. and Diaconis, P. (2010). A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics* **6**, 489-522.

Blondel, V. D., Guillaume, J. L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **10**, P10008.

Bollobás, B. (1980). A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics* 1, 311-316.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z. and Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* **20**, 172-188.

Cafieri, S., Hansen, P. and Liberti, L. (2010). Loops and multiple edges in modularity maximization of networks. *Physical Review E* **81**, 046102.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

Chatterjee, S., Diaconis, P. and Sly, A. (2011). Random graphs with a given degree sequence. *Annals of Applied Probability* **21**, 1400-1435.

Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* **22**, 2283-2290.

Choi, D. S., Wolfe, P. J. and Airoldi, E. M.(2011). Confidence sets for network structure. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **4**, 461-469.

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

Chung, F. and Lu, L. (2002). Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics* **6**, 125-145.

Dourisboure, Y., Geraci, F. and Pellegrini, M. (2007). Extraction and classification of dense communities in the Web. *Proceedings of the 16th International Conference on World Wide Web*, 461-470.

Fienberg, S. E., Meyer, M. M. and Wasserman, S. S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association* **80**, 51-67.

Flake, G. W., Lawrence, S. and Giles, C. L. (2000). Efficient identification of Web communities. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 150-160.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports* **428**, 75-174.

Guimera, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Physical Review E* **70**, 025101.

Handcock, M. S., Raftery, A. E. and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A* **170**, 301-354.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association* **76**, 33-50.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

Jin, J. (2015). Fast community detection by SCORE. *The Annals of Statistics* **43**, 57-89.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* **83**, 016107.

Lehmann, E. L. (1986). *Testing Statistical Hypothesis*, 2nd ed. Wiley, New York.

Massen, C. and Doye, J. (2005). Identifying communities within energy landscapes. *Physical Review E* **71**, 046101.

McKay, B. D. (1985). Asymptotics for symmetric 0-1 matrices with prescribed row sums. *Ars Combinatoria* **19A**, 15-26.

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

Newman, M. E. J. (2006a). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 8577-8582.

Newman, M. E. J. (2006b). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* **74**, 035104.

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

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic block structures. *Journal of the American Statistical Association* **96**, 1077-1087.

Park, J. and Newman, M. E. J. (2004). Statistical mechanics of networks. *Physical Review E* **70**, 066117.

Peng, L. and Carvalho, L. (2013). Bayesian degree-corrected stochastic block models for community detection. Preprint arXiv:1309.4796.

Perry, P. O. and Wolfe, P. J. (2012). Null models for network data. Preprint arXiv:1201.5871.

Reddy, K. P., Kitsuregawa, M., Sreekanth, P. and Rao, S. S. (2002). A graph based approach to extract a neighborhood customer community for collaborative filtering. *Proceedings of the Second International Workshop on Databases in Networked Information Systems*, 188-200.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0040

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E* **74**, 016110.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 888-905.

Snijders, T. A. B. and Nowicki, K. (1997). Estimation and prediction for stochastic block models for graphs with latent block structure. *Journal of Classification* **14**, 75-100.

Wang, G., Shen, Y. and Ouyang, M. (2008). A vector partitioning approach to detecting community structure in complex networks. *Computers and Mathematics with Applications* **55**, 2746-2752.

Yang, J., McAuley, J. and Leskovec, J. (2013). Community detection in networks with node attributes. *Proceedings of the IEEE International Conference on Data Mining*, 1151-1156.

Zhang, J. and Chen, Y. (2013). Sampling for conditional inference on network data. *Journal of the American Statistical Association* **108**, 1295-1307.

Zhao, Y., Levina, E. and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics* **40**, 2266-2292.

Department of Management Science, University of Miami, Coral Gables, FL 33124-6544, USA.

E-mail: ezhang@bus.miami.edu

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: yuguo@illinois.edu