The Report Commitee for Lingjia Zhang
Certifies that this is the approved version of the following report

Community Detection in Network Analysis: A Survey

APPROVED BY

SUPERVISING COMMITTEE:

Lizhen Lin, Supervisor

Timothy Keitt

Community Detection in Network Analysis: A Survey

by

Lingjia Zhang, B.S.

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN STATISTICS

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2016

Community Detection in Network Analysis: A Survey

Lingjia Zhang, M.S.Stat.
The University of Texas at Austin, 2016

Supervisor: Lizhen Lin

The existence of community structures in networks is not unusual, including in the domains of sociology, biology, and business, etc. The characteristic of the community structure is that nodes of the same community are highly similar while on the contrary, nodes across communities present low similarity. In academia, there is a surge in research efforts on community detection in network analysis, especially in developing statistically sound methodologies for exploring, modeling, and interpreting these kind of structures and relationships. This survey paper aims to provide a brief review of current applicable statistical methodologies and approaches in a comparative manner along with metrics for evaluating graph clustering results and application using *R*. At the end, we provide promising future research directions.

Table of Contents

Chapter 1

Introduction

A network or a graph is a collection of points joined by lines, and we call these points nodes and the lines edges. Nodes in the network may represent individuals, organizations, or some other kind of units of study; edges correspond to types of links, relationships, or interactions between the nodes. Network analysis has become one of most popular modern research topics. Caldarelli and Vespignani [4] brought up the concept of a complex network, which is "a system composed of interconnected parts which, as a whole, exhibits one or more properties [...] not obvious from the properties of the individual parts." Essentially, network is a visual way of analyzing and exploring different relationships. When we analyze the whole network, we can learn new insights that we would not necessarily know just by looking at individual piece of the network.

Networks are everywhere. There are social networks, such as sexual networks, criminal networks, and interaction networks over online social networking communities (e.g., *Facebook*, *Twitter*, and *LinkedIn* are recent phenomenon). There are also biological networks, including protein-protein interaction networks, neuronal networks, food webs, and species interaction net-

works. Business networks, such as financial networks, supply chains, and retail networks are also prevalent. Many of them are inhomogeneous, consisting of not only massive nodes but distinct communities. Clustering, or community structure detection, divides network nodes into groups within which the network connections are dense (i.e., there are more edges between nodes), but between which they are sparser (i.e., fewer edges) [22]. Clusters are present in networks, for example, as prospective groups and their friendships on social media, or as modules of functionally associated proteins in biological networks, or as a neighborhood community of customers with similar purchasing preferences, and much more. As such, communities or clusters of highly connected nodes form an essential feature in the structure of networks, and the identification of these communities is essential in answering important questions in a variety of fields.

Due to the extent and the diversity of contexts in which networks appear, community detection in network analysis has become a both crucial and interdisciplinary topic. However, finding clusters or detecting communities in networks is a challenging task in a wide range of domains, especially for directed networks. For instance, a directed graph is characterized by asymmetrical matrices (e.g., *adjacency matrix*, *Laplacian matrix*, which we introduce in Section 2.1), so spectral analysis is much more complex. Only a few methods can be easily extended from the undirected to the directed case [9]. Further, extracting clusters in networks is algorithmically difficult because it is computationally intractable to search over all possible clustering [27]. This ne-

cessitates the development of statistically sound methodologies for exploring, modeling, and interpreting these relationship in networks.

The main goal of this survey paper is to review some of the major statistical methods and algorithms proposed in the research communities for the problem of community detection in static networks in a comparative manner. Some of them are new methods while others extend approaches that have been previously applied on network analysis. Several of the statistical models and methodologies we have summarized are shown to perform very well in detecting community structures on a variety of real-world networks like the ones presented above. To name a few, latent space approaches are applied in social science where it studies marriage and business relations [14]. Stochastic blockmodels are used in the analysis of protein-protein interactions where blocks may correspond to stable protein complexes [11]. Modularity optimization is applied on marine sciences where it successfully detects the main two-way division of the dolphin social network [20].

The rest of the paper is organized as follows. In Chapter 2, we provide the basic terminology and background used throughout this survey. Then in Chapter 3, we present main clustering approaches developed for both undirected and directed networks. We also present an empirical comparison of the main methods that have been reviewed throughout this paper. In Chapter 4, we introduce R's package "igraph" for community detection in network analysis and present a data example by using one of its built-in algorithms. Finally in Chapter 5, we draw conclusions from this overview by summarizing this

survey and inferring future research directions. For an extensive review, see Goldenberg et al. [11] which provides a review of the literature of statistical modeling and analysis of networks including discussions of both static and dynamic network modeling.

Chapter 2

Basic Terminology and the Quality Measurement

In this chapter we provide basic terminology and background used in network analysis. We firstly introduce the notation and terminology, including some basic graph theory and linear algebraic concepts. Then we describe briefly the major metrics used to quantify the quality of a community in networks.

2.1 Basic Network Notation and Terminology

For the purpose of describing various methods and algorithms, we introduce the following notations. A *graph* or a *network* G is often defined in terms of nodes and edges: $G \equiv G(N, E)$, where N represents the node set and E is the edge set. In computer science, networks contain nodes and edges; while in social sciences, the corresponding terminology is actors and ties [11]. In this review, we use these terms interchangeably.

A basic property of the nodes in a graph is their degree, that is, the number of edges that connect to this node. The *degree matrix* is defined as the diagonal and positive $N \times N$ matrix D, with the degree of each node in

the main diagonal and zeros outside the main diagonal.

Edges may be undirected as in the Erdös-Rényi-Gilbert model, or directed as in the Holland and Leinhardt's p_1 model. In a directed graph, the edge set E contains an ordered pair of nodes (i, j) if there is an edge, or relationship, from the node i to node j; in an undirected one, if the edge set contains (i, j), then (j, i) as well. The edge set E can be represented by the *adjacency matrix* Y of size $N \times N$ with binary elements in a setting where we only concern about the presence or absence of edges: $Y \in \{0, 1\}^{N \times N}$, thus $G \equiv G(N, Y)$ and

$$Y_{i,j} = \begin{cases} 1, & \text{if (i, j) is in the edge set} \\ 0, & \text{otherwise} \end{cases}. \tag{2.1}$$

As such, for undirected relations where $Y_{j,i} = Y_{i,j}$, the adjacency matrix is symmetric; while in a directed network, Y is not necessarily symmetric.

Other than using an adjacency matrix Y represent a graph G, we can also associate each graph with its *Laplacian matrix* that is defined using linear algebraic concepts. Given a simple graph G with n nodes, its Laplacian matrix $L_{N \times N}$ is defined as

$$L = D - Y, \tag{2.2}$$

where D is the degree matrix and Y is the adjacency matrix of the graph.

The *symmetric normalized Laplacian matrix* is defined as:

$$L^{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} Y D^{-1/2}, \tag{2.3}$$

where L is the (unnormalized) Laplacian, Y is the adjacency matrix, and D is the degree matrix.

2.2 Quality Measure by Modularity

In practical situations, communities within a network are often not known beforehand. This raises the following question: how to measure whether the community structure found by the algorithm is a good one? Modularity, which is proposed by Newman and Girvan [22], is one of the most popular and widely used metrics to evaluate the quality of network's division into communities. Informally, the *modularity* Q of each possible partition will be:

$$Q = (\text{fraction of edges within communities}) - (\text{expected fraction of edges}). \tag{2.4}$$

More precisely, consider a particular division of a network into k communities. Define a $k \times k$ symmetric matrix $\mathbf{e}$ whose element e_{ij} is the fraction of all edges in the network that link nodes in community i to nodes in community j. Then

$$Q = \sum_i (e_{ii} - a_i^2) = Tr(\mathbf{e}) - \parallel \mathbf{e}^2 \parallel, \tag{2.5}$$

where a_i is the row (or column) sums $\sum_j e_{ij}$ which represents the fraction of edges that connects to nodes in community i, the trace of matrix $\mathbf{e}$, $Tr(\mathbf{e}) = \sum_i e_{ii}$, gives the fraction of edges in the network that connect nodes in the same community, and $\parallel \mathbf{x} \parallel$ indicates the sum of elements of the matrix $\mathbf{x}$.

Clearly, larger positive values of modularity indicate better division into communities since there are more edges within communities than one would expect if edges were placed in random. If the number of within-community

edges is no better than random, $Q = 0$; while Q is approaching to 1, which is the maximum, it indicates networks with strong community structure [22].

Other than being used as quality measure for a specific network partition, modularity can also be used for detecting community structures in networks [20]. This procedure is described in more detail in Section 3.5.

Chapter 3

Existing Methodologies

This chapter firstly summarizes some of original work that has been done on analyzing network models: the research originates with the Erdös-Rényi-Gilbert random graph model ([7], [10]), and the p_1 model of Holland and Leinhardt [17] in some sense generalizes the Erdös-Renyi-Gilbert model. Section 3.2 discusses the stochastic blockmodels, which is a special version of p_1 model that could be used to describe a random graph model with predefined blocks. Section 3.3 summarizes latent space approaches for social network analysis by Hoff et al. [14], followed by Handcock et al. [12]'s latent position cluster model, which is an application of the latent space model for clustering. Section 3.4 describes spectral clustering and the high-dimensional stochastic blockmodel proposed by Rohe et al. [27]. Finally in Section 3.5, we describe modularity's usage for detecting community structure in networks.

3.1 The Erdös-Rényi-Gilbert Random Graph Model and the p_1 Model

For a binary graph with conditionally independent edges, each edge outcome $y_{i,j}$ is a dichotomous variable indicating the presence ($y_{i,j} = 1$) or absence ($y_{i,j} = 0$) of some relation or edge. It can be expressed as a Bernoulli

binary random variable with probability of presence π_{ij}.The simplest case of this class of network probability models was introduced contemporaneously by Erdös and Rényi [7] and Gilbert[10]: known as the Erdös-Rényi-Gilbert random graph model. This basic model describes an undirected graph involving N nodes and a fixed number of edges E, chosen randomly from $m = \binom{N}{2}$ possible edges in the graph G. All edges essentially have the same probability $\pi_{ij} = p$ of presence and are independent from one another, thus the binomial likelihood of the Erdös-Rényi-Gilbert random graph model $G(N, p)$ is

$$l(G(N,p) \text{ has } E \text{ edges } |p) = p^E(1-p)^{m-E}, \tag{3.1}$$

or, equivalently in terms of the $N \times N$ adjacency matrix Y

$$l(Y|p) = \prod_{i \neq j} p^{Y_{ij}}(1-p)^{1-Y_{ij}}. \tag{3.2}$$

Empirically there are few observed networks with such simple structure as in the Erdös-Rényi-Gilbert random graph model. This has led to the p_1 model of Holland and Leinhardt [17], which began with a directed version of the Erdös-Renyi-Gilbert random graph model and proposed that three parameters affect the outcome of a dyad with directed edges: 1). "reciprocity" ρ, that is, the tendency of $y_{i,j} = y_{j,i}$; 2). "gregariousness" α of an actor, that is, how likely one is to have outgoing ties; 3). the "popularity" β of an actor, that is, how likely one is to have incoming ties.

Let $P(0,0)$ be the probability for the absence of an edge between i and j, $P_{ij}(1,0)$ the probability of i linking to j ("1" indicates the outgoing node of

the edge), $P_{ij}(1,1)$ the probability of i linking to j and j linking to i. Given a parameter for the overall density of edges θ, the p_1 model posits the following probabilities [17]:

$$\begin{aligned}
\log P_{ij}(0,0) &= \lambda_{ij}, \\
\log P_{ij}(1,0) &= \lambda_{ij} + \alpha_i + \beta_j + \theta, \\
\log P_{ij}(0,1) &= \lambda_{ij} + \alpha_j + \beta_i + \theta, \\
\log P_{ij}(1,1) &= \lambda_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + 2\theta + \rho_{ij},
\end{aligned} \tag{3.3}$$

where λ_{ij} is a normalized constant to ensure that the total probabilities for each dyad (i,j) add up to 1.

The form of the joint likelihood for the $p1$ model is in exponential family form,

$$\log P(Y = y) \propto \theta y_{++} + \sum_i \alpha_i y_{i+} + \sum_j \beta_j y_{+j} + \rho \sum_{ij} y_{i,j} y_{j,i}, \tag{3.4}$$

where a "+" denotes summing over the corresponding subscript. The minimum sufficient statistics are the in-degree (i.e., y_{i+}) and out-degree (i.e., y_{+j}) for each node and the number of dyads with reciprocated edges (i.e., $\sum_{ij} y_{ij} y_{ji}$). Holland and Leinhardt [17] presented an iterative proportional fitting method for maximum likelihood estimation for this model.

A major problem with the Erdös-Rényi-Gilbert random graph model and the p_1 model is that the complexities involved in assessing goodness-of-fit procedures for the model [17]. Also, these models are restrictive as they assume the $\binom{N}{2}$ dyads (y_{ij}, y_{ji}) to be independent [14].

3.2 Stochastic Blockmodels

Community detection, in another sense, is to search for an optimal partition of the nodes in a network into groups or blocks. This is known as blockmodeling. Many researchers have extended the p_1 model to blockmodels. For example, within the framework of p_1 model and its exponential family generalizations, Nowicki and Snijders ([30], [23]) developed models for a restricted version of the blockmodel in which group membership is not observed. Blockmodeling is becoming a common approach in network analysis to decompose a graph.

The stochastic blockmodel, introduced by Holland et al. [16], is a special version of p_1 model that could be used to describe a random graph model with predefined blocks. This model tends to produce graphs containing communities characterized by being connected with one another with particular edge densities. For example, edges may be more common within communities than between communities. It is also an example of the more general latent space model of a random network by Hoff et al. [14] which we will describe in more detail in Section 3.3.

The idea that nodes heavily interconnected should form a block or community forms the basic of stochastic blockmodel. The nodes are reordered to display the blocks down the diagonal of the adjacency matrix Y. Further, the connections between nodes in different blocks appear in much sparser off-diagonal blocks. In model-based approaches, the partition of the nodes maximizes the likelihood function linked to the model, whereas most algorithmic

solutions maximize ad hoc criteria related to the "density" of links within and between blocks [11].

An important assumption of stochastic blockmodel relies on the intuitive notion of *structural equivalence*; that is, the probability of connectivity between (i_1, j_1) is the same as that of (i_2, j_2) if nodes i_1 and j_1 are in the same respective latent classes as i_2 and j_2 [14]. As such, it is useful in the analysis of social relations where blocks may correspond to social factions, as well as in the analysis of protein-protein interactions where blocks may correspond to stable protein complexes.

The stochastic blockmodel is characterized by the fact that each node belongs to one of multiple blocks and the probability of a relationship between two nodes depends only on the block memberships of the two nodes. If the probability of an edge between two nodes in the same block is larger than the probability of an edge between two nodes in different blocks, then the blocks produce communities in the random networks generated from the model [27].

However, stochastic blockmodel is restrictive, as they only fit well when *stochastic equivalence* for clusters of individuals holds but not when many actors fall between clusters, or when relations are transitive yet there is no strong clustering[14].

3.3 Latent Space Approaches

Latent variable model is generally used when some variables are not directly observable but are accounting for the unobserved heterogeneity between subjects. Hoff et al. [14] proposed a class of latent space models in the context of social network analysis in 2002. The intuition at the core of their models is that each actor is assumed to have a latent position, z_i, in a social space and since the positions are unknown, the social space is a latent variable.

In their methods, Hoff et al. [14] modeled the positions as belonging to a low-dimensional Euclidean space. As such, the existence of an edge in the adjacency matrix, $Y_{i,j} = 1$, is determined by the distance, $d(z_i, z_j)$, among the corresponding pair of actors in the low-dimensional space, and also by the values of a number of covariates observed on each actor individually if further covariate information is available. Therefore, the model derives the probability from the distance between latent representations. That is, actors are likely to be in a relationship if their latent representations are close according to the Euclidean distance. Assuming that the presence or absence of a tie between two individuals is independent of all other ties in the system, given the unobserved positions in social space of the two individuals, the conditional probability model for the adjacency matrix Y is

$$P(Y|Z, X, \Theta) = \prod_{i \neq j} P(y_{i,j}|z_i, z_j, x_{i,j}, \Theta), \tag{3.5}$$

where X are observed covariates, Θ are parameters, and Z are the positions of actors in the low-dimensional latent space. Each relationship $Y_{i,j}$

is a sample from a Bernoulli distribution whose natural parameter depends on $z_i, zj, x_{i,j}, \Theta$. In their model, Hoff et al. [14] generated the paired observations $Y_{i,j}$ starting from the corresponding pair of actors representations (z_i, z_j) through a distance model, pair-specific and vector-valued covariates $x_{i,j}$ for dyad (i, j), and parameters $\Theta = (\alpha, \beta)$.

A convenient parameterization of $P(Y|Z, X, \Theta)$ is then the log-odds ratio:

$$\log \frac{P(Y_{i,j} = 1)}{1 - P(Y_{i,j} = 1)} = \alpha + \beta' x_{i,j} - |z_i - z_j| = \eta_{i,j}, \tag{3.6}$$

and the corresponding log-likelihood function is

$$\log P(Y|\eta) = \sum_{i \neq j} \{\eta_{i,j} y_{i,j} - \log(1 + e^{\eta_{i,j}})\}. \tag{3.7}$$

The log-likelihood function, which is equivalent to the likelohood function of nonlinear logistic regression models, can be maximized to obtain maximum-likelihood estimates. Another feasible approach is based on Bayesian inference. Given prior information on α, β, and Z, use Gibbs sampling to sample from the posterior distribution of α, β, and Z. However, distances between a set of points in Euclidean space are invariant under rotation, reflection, and translation. Hoff et al. [14] addressed this problem by using a "Procrustean" transformation to rotate and reflect these posterior draws to be as close as possible to a reference configuration.

Hoff et al. [14] proposed a model that has several advantages over the previous described models. In addition to improving on model fit, modeling

the positions as belonging to a low-dimensional Euclidean space provided a model-based spatial representation of network relationships. Also, it allowed statistical uncertainty in the social space to be quantified and graphically represented. The model is flexible and able to deal with missing data. Finally, the model is inherently transitive. The latent space model has been recently extended in a number of directions to include treatment of transitivity, homophily on actor-specific attributes, clustering, and heteroheneity of nodes [11]. For future works, it may be desirable to allow for further dependence in the model [14] and scalability issues remain to be addressed before larger networks can be analyzed [11].

3.3.1 The Latent Position Cluster Model

The latent space model has been recently extended in a number of directions ([15], [13], [29]). Recall that in latent space approaches, each actor has a latent position in a low-dimensional Euclidean space with potential further information on covariates. Handcock et al. [12] extended this approach through a combination of latent space models with model-based clustering, thus proposed the latent position cluster model.

To allow joint inference on latent positions and clustering, the latent position cluster model combines the original latent space model with a finite mixture of Gaussians approach to clustering. That is to say, the authors assumed that the z_is are drawn from a finite mixture of G multivariate normal distributions, each representing a different group of actors. Thus

$$
\begin{aligned}
P(Y|Z, X, \Theta) &= \prod_{i \neq j} P(y_{i,j}|z_i, z_j, x_{i,j}, \Theta) \\
z_i &\sim \sum_{g=1}^{G} \lambda_g \text{MVN}_d(\mu_g, \sigma_g^2 I_d)
\end{aligned}
\tag{3.8}
$$

where λ_g is the probability that an actor belongs to the gth group so that $\sum_{g=1}^{G} \lambda_g = 1$, and I_d is the $d \times d$ identity matrix.

For estimating the latent positions and the model parameters, Handcock et al. [12] found that a fully Bayesian method that uses Markov chan Monte Carlo (MCMC) sampling performs better than using maximum likelihood estimation procedure. The latent position cluster model captures transitivity, homophily on attributes, and clustering simultaneously. As a result, it can be viewed as not only a stochastic blockmodel with transitivity within blocks and homophily on attributes, but also a generalization of latent class models to allow heterogeneity of structure within the classes [12].

3.4 Spectral Clustering and the High-dimensional Stochastic Blockmodel

Spectral clustering is a nonparametric algorithm initialed by the work of Donatha and Hoffman [6] and Fiedler [8], and can identify the connected components in a graph (if there are any) by making use of the eigenvalues of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. In the context of network analysis, spectral clustering is a popular and computationally feasible method to discover com-

munities in undirected networks. The similarity matrix, which is provided as an input, is the adjacency matrix Y that we introduced in the equation 2.1.

Define the *symmetric normalized graph Laplacian* L as

$$L = D^{-1/2} Y D^{-1/2} \tag{3.9}$$

where D is the degree matrix, and

$$D_{ii} = \sum_k Y_{i,k}. \tag{3.10}$$

Note that this definition does not contradict with that in the equation 2.3. For spectral clustering, the difference is immaterial because both definitions have the same eigenvectors.

Rohe et al. [27] studied the performance of spectral clustering, a non-parametric method, on a parametric task of partitioning graphs into blocks in the stochastic blockmodel. Basically, they bound the number of nodes "misclustered" for networks generated from the stochastic blockmodel using spectral clustering. Specifically, the algorithm for k many clusters is defined in the following way:

1. Take the symmetric adjacency matrix $Y \in \{0,1\}^{n \times n}$ as input.

2. Find the eigenvectors $X_1, \ldots, X_k \in R^n$ corresponding to the k eigenvalues of L that are largest in absolute value. Since L is symmetric, choose these eigenvectors to be orthogonal. By putting the eigenvectors into the columns, form the matrix $X = [X_1, \ldots, X_k] \in R^{n \times k}$.

3. Treating each of the n rows in X as a point in R^k, run k-means with k clusters. This creates k nonoverlapping sets $A_1, \ldots, A_k$ whose union is $1, \ldots, n$.

4. Output: $A_1, \ldots, A_k$. This means that node i is assigned to cluster g if the i-th row of X is assigned to A_g in step 3.

In their paper, Rohe et al. [27] showed two main results. The first main result is that under the more general latent space model, the top k eigenvectors of the normalized graph Laplacian L are consistent, in the sense that they asymptotically converge to the eigenvectors of a "population" normalized graph Laplacian as the number of nodes n grows to infinity. They also provided guarantees on the performance of a spectral clustering algorithm based on the normalized Laplacian.

The second main result, by assuming the probability of an edge between two nodes is p, if $p \neq 0$ and $k = O(n^{1/4}/logn)$, then the number of nodes that might be misclustered by running k-means is:

$$|M| = o(k^3(\log n)^2). \tag{3.11}$$

This proves that if the minimum expected degree grows fast enough and the smallest nonzero eigenvalue of the population normalized graph Laplacian shrinks slowly enough, then the proportion of nodes that are misclustered by spectral clustering converges to zero:

$$\frac{|M|}{n} = o(n^{-1/4}). \tag{3.12}$$

The asymptotic framework proposed by Rohe et al. [27] allows the number of clusters in the stochastic blockmodel to grow with the number of nodes, hence it makes the problem one of "high-dimensional" learning.

Rohe et al. [27]'s spectral clustering provides a computationally appealing alternative to maximum likelihood fitting in practice. However, they only considered graphs where the expected degrees of nodes in the same cluster are equal. Further, studying spectral clustering under more realistic degree distributions is an area for future research.

3.5 Modularity Optimization

As we discussed in Section 2.2, modularity is one of the most popular and widely used metrics to evaluate the quality of network's division into communities. In this section, we described its function for extracting the community structure in networks.

Newman [20] has proposed an approach to the discovery of community structure based on the modularity Q defined in the equation 2.5 [20]. Recall that the modularity is defined as

$$Q = \sum_i (e_{ii} - a_i^2), \tag{3.13}$$

where matrix $\mathbf{e}$ has element e_{ij} that is the fraction of all edges in the network that link nodes in community i to nodes in community j, a_i represents the fraction of edges that connect to nodes in community i and $a_i = \sum_j e_{ij}$.

Since we learned from Section 2.2 that a high value of Q represents a

good community division, Newman [20] proposed that one ought to be able to find the best communities in a network by optimizing Q over all possible divisions. Specifically, the author offered a greedy algorithm that starts with each node in a separate community on its own, and repeatedly join communities together in pairs, choosing at each step the join that gives the greatest increase (or smallest decrease) in Q. Thus, the whole procedure can be represented as a "dendrogram," a tree that shows the order of the joins. The optimal cross-section of the "dendrogram" is found by looking for the maximal value of Q.

The main advantage of Newman [20]'s modularity optimization algorithm is its speed, which allows the analysis of large networks where communities are substantial in size and composed of many individuals. In addition, it provided a useful tool for visualizing and understanding the structure of theses networks, whose daunting size has hitherto made many of their structural properties obscure [20]. However, in more recent Bickel and Chen's study [1], the authors imply that using modularity scores are (asymptotically) biased: it leads to incorrect community structure discovery even in the favorable case of large networks.

Chapter 4

Application in R

In this chapter, we introduce the *R*'s package "igraph" for community detection in network analysis. Then we present how one could conduct this type of analysis using one of its various built-in algorithms on a data example.

4.1 *R* package - igraph

There are several ways to do community partitioning of graphs using very different packages. The most popular package is "igraph". Not only that 95% of what one will need in network analysis is available in "igraph", but that the libraries are written in *C* and therefore are fast.

In the "igraph" package there are a few already implemented community detection algorithms for clustering, partitioning, and segmenting a network, including some we have introduced in Chapter 3:

- **edge.betweenness.community** [22]: a divisive algorithm where at each step the edge with the highest betweenness is removed from the graph. For each division one can compute the modularity of the graph and then choose to cut the "dendrogram" where the process gives the highest value of modularity.

- **fastgreedy.community** [5]: the algorithm is agglomerative and at each step the merge is decided by the optimization of modularity that it produces are the result of the merge.

- **label.propagation.community** [25]: a nearly linear time algorithm by labeling the vertices with unique labels and then updating the labels by majority voting in the neighborhood of the vertex.

- **leading.eigenvector.community** [21]: tries to find densely connected subgraphs in a graph by calculating the leading non-negative eigenvector of the modularity matrix of the graph.

- **multilevel.community** [2] (the Louvain method) implements the multilevel modularity optimization algorithm which is based on the modularity mearsure and a hierarchical approach.

- **optimal.community** [3] calculates the optimal community structure of a graph by maximizing the modularity measure over all possible partitions.

- **spinglass.community** [26] uses as spin-glass model and simulated annealing to find the communites inside a network.

- **walktrap.community** [24] finds densely connected subgraphs by performing random walks. The idea is that random walks will tend to stay inside communities instead of jumping to other communities.

- **infomap.community** [28] finds community structure that minimizes the expeted description length of a random walker trajectory.

All of these methods return a "communities" object, which one can then use to explore, plot, and compute metrics. For documentation on how to use "igraph", a manual is available at `http://igraph.org/r/doc/aaa-igraph-package.html`.

4.2 Data Examples

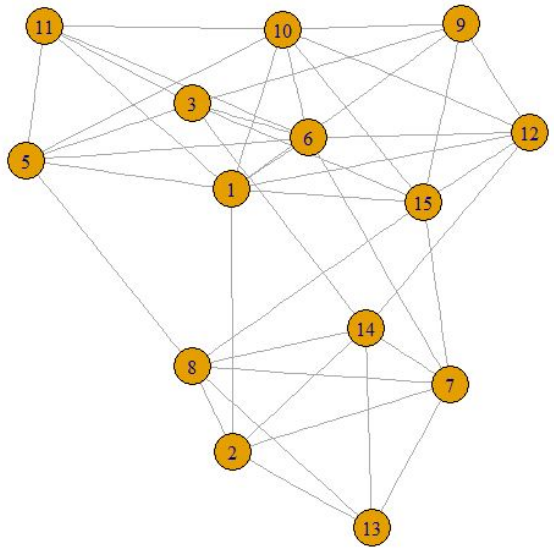
Here we use a data example of student networks from the lab source of the Social Network Analysis Group at Stanford University [18] to illustrate how those community detection algorithms in the *R* package "igraph" can be used.

The task is to identify friendship groups or communities and to discern the best fitting community structure in an undirected network as shown in the Figure 4.1. Note that for clarity and simplicity, we removed isolated vertices.

There are many different functions as we have shown in previous section that can be used in the package "igraph". We chose to use the edge-betweenness algorithm from Newman [22].

As discussed above, the idea of the edge-betweenness algorithm is that it is likely that edges connecting separate cluster have high edge-betweenness, as all the shortest paths from one cluster to another must traverse through them. So iteratively remove the edge with the highest betweenness from the

Figure 4.1: Plot for friend layout



graph, we will get a hierarchical map of the communities in the graph, called a "dendrogram" (see Figure 4.2). The leafs of the tree are the individual vertices and the root of the tree represents the whole graph. As such, we can tell from the "dendrogram" that there are three clusters in this network.

Figure 4.3 shows all modularities for each merge/division. From that, we can then choose to cut the "dendrogram" where the process gives the highest value of modularity. Figure 4.4 shows the colored nodes according to their membership after the clustering process.

Figure 4.2: Visualization as a dendrogram

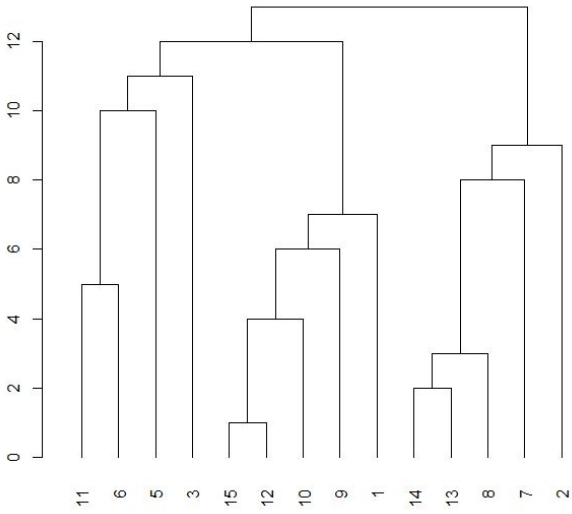


Figure 4.3: Modularity for each merge

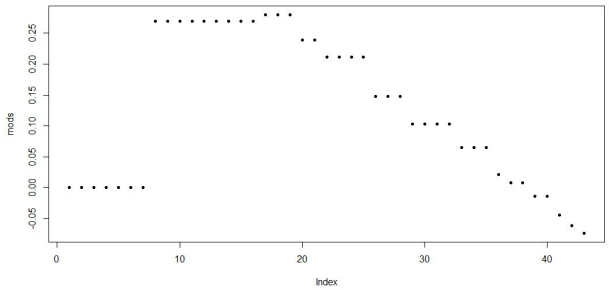
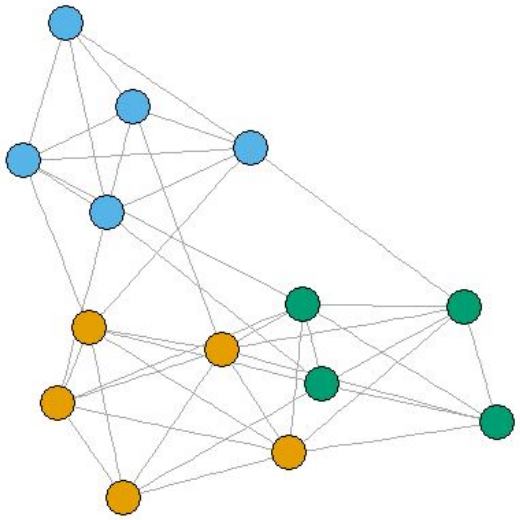


Figure 4.4: Colored nodes according to membership



Chapter 5

Conclusion and Future Directions

In this survey we have reviewed several statistical methodologies for exploring, modeling, and interpreting community structures in network data. We firstly introduced some basic notation and concepts in graph theory and linear algebra (Section 2.1), followed by describing the modularity quality measure (Section 2.2). Then we discussed some of the earliest works in this field, such as the Erdös-Rényi-Gilbert random graph model and the p_1 model (Section 3.1). However, as we have pointed out, these approaches have shortcomings as far as the complexities involved in assessing goodness-of-fit procedures and also concerning on the analysis of large real-world networks. This leads to various new models being developed that are flexible enough to apply on general network structures in the last few years. We have presented the popular stochastic blockmodels (Section 3.2). We have also described in detail several major approaches that is based on and extended from earlier works, including the latent space approaches(Section 3.3), the latent position cluster model (Section 3.3.1), spectral clustering algorithms (Section 3.4), as well as modularity optimization (Section 3.5). Moreover, we introduced the most popular R's package for community detection in network analysis - "igraph" and presented a data example by using one of its built-in algorithms (Chapter 4).

The main goal of this survey report is to review these major statistical methods and algorithms proposed so far for the problem of community detection in networks and outlined each method's strengths and weaknesses. But there are many issues that are remained to be solved, such as on network visualization, computability, and assessing goodness of fit. Therefore, we feel that there is still scope for developing systematic ways to visualize community structures in networks in the areas of inference and dynamic modeling. For example, creating or extending an existing model (e.g., bayesian models and placing its prior on partitions) in a way that provides inference mechanisms which can infer parameters of large scale networks would be a great breakthrough to the statistical network modeling community.

Bibliography

[1] Peter J Bickel and Aiyou Chen, *A nonparametric view of network models and newman–girvan and other modularities*, Proceedings of the National Academy of Sciences **106** (2009), no. 50, 21068–21073.

[2] VD Blondel, JL Guillaume, R Lambiotte, and E Lefebvre, *Fast unfolding of community hierarchies in large network, 2008*, J. Stat. Mech. P **1008**.

[3] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner, *On modularity clustering*, Knowledge and Data Engineering, IEEE Transactions on **20** (2008), no. 2, 172–188.

[4] Guido Caldarelli and Alessandro Vespignani, *Large scale structure and dynamics of complex networks*, World Scientific, 2007.

[5] Aaron Clauset, Mark EJ Newman, and Cristopher Moore, *Finding community structure in very large networks*, Physical review E **70** (2004), no. 6, 066111.

[6] William E Donath and Alan J Hoffman, *Lower bounds for the partitioning of graphs*, IBM Journal of Research and Development **17** (1973), no. 5, 420–425.

[7] Paul Erd6s and A Rényi, *On the evolution of random graphs*, Publ. Math. Inst. Hungar. Acad. Sci **5** (1960), 17–61.

[8] Miroslav Fiedler, *Algebraic connectivity of graphs*, Czechoslovak mathematical journal **23** (1973), no. 2, 298–305.

[9] Santo Fortunato, *Community detection in graphs*, Physics reports **486** (2010), no. 3, 75–174.

[10] Edgar N Gilbert, *Random graphs*, The Annals of Mathematical Statistics **30** (1959), no. 4, 1141–1144.

[11] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi, *A survey of statistical network models*, Foundations and Trends® in Machine Learning **2** (2010), no. 2, 129–233.

[12] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum, *Model-based clustering for social networks*, Journal of the Royal Statistical Society: Series A (Statistics in Society) **170** (2007), no. 2, 301–354.

[13] Peter D Hoff, *Bilinear mixed-effects models for dyadic data*, Journal of the american Statistical association **100** (2005), no. 469, 286–295.

[14] Peter D Hoff, Adrian E Raftery, and Mark S Handcock, *Latent space approaches to social network analysis*, Journal of the american Statistical association **97** (2002), no. 460, 1090–1098.

[15] Peter D Hoff and Michael D Ward, *Modeling dependencies in international relations networks*, Political Analysis **12** (2004), no. 2, 160–175.

[16] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt, *Stochastic blockmodels: First steps*, Social networks **5** (1983), no. 2, 109–137.

[17] Paul W Holland and Samuel Leinhardt, *An exponential family of probability distributions for directed graphs*, Journal of the american Statistical association **76** (1981), no. 373, 33–50.

[18] Solomon Messing Mike Nowak McFarland, Daniel and Sean Westwood, *Social network analysis labs in r*, 2010 (accessed April 25, 2016).

[19] Mark EJ Newman, *Detecting community structure in networks*, The European Physical Journal B-Condensed Matter and Complex Systems **38** (2004), no. 2, 321–330.

[20] ———, *Fast algorithm for detecting community structure in networks*, Physical review E **69** (2004), no. 6, 066133.

[21] ———, *Finding community structure in networks using the eigenvectors of matrices*, Physical review E **74** (2006), no. 3, 036104.

[22] Mark EJ Newman and Michelle Girvan, *Finding and evaluating community structure in networks*, Physical review E **69** (2004), no. 2, 026113.

[23] Krzysztof Nowicki and Tom A B Snijders, *Estimation and prediction for stochastic blockstructures*, Journal of the American Statistical Association **96** (2001), no. 455, 1077–1087.

[24] Pascal Pons and Matthieu Latapy, *Computing communities in large networks using random walks*, Computer and Information Sciences-ISCIS 2005, Springer, 2005, pp. 284–293.

[25] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara, *Near linear time algorithm to detect community structures in large-scale networks*, Physical Review E **76** (2007), no. 3, 036106.

[26] Jörg Reichardt and Stefan Bornholdt, *Statistical mechanics of community detection*, Physical Review E **74** (2006), no. 1, 016110.

[27] Karl Rohe, Sourav Chatterjee, and Bin Yu, *Spectral clustering and the high-dimensional stochastic blockmodel*, The Annals of Statistics (2011), 1878–1915.

[28] Martin Rosvall and Carl T Bergstrom, *Maps of random walks on complex networks reveal community structure*, Proceedings of the National Academy of Sciences **105** (2008), no. 4, 1118–1123.

[29] Michael Schweinberger and Tom AB Snijders, *Settings in social networks: A measurement model*, Sociological Methodology **33** (2003), no. 1, 307–341.

[30] Tom AB Snijders and Krzysztof Nowicki, *Estimation and prediction for stochastic blockmodels for graphs with latent block structure*, Journal of classification **14** (1997), no. 1, 75–100.

Vita

Lingjia Zhang was born in Chengdu, China in 1991. She earned her Bachelor of Science degree in Mathematics and a Minor in Management Information Systems at The University of Texas at Arlington. Lingjia interned with Alcon in Fort Worth and Hewlett-Packard (HP) in Houston as a Quality Analyst during and after college. After completing her internship at HP, she entered the University of Texas at Austin and started graduate studies in Statistics on August, 2014. Lingjia's research interests include time series and statistical methods for social network analysis.

This report was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.