THE TITLE OF YOUR THESIS:
USE LINE BREAKS IF NEEDED

Your M. Name

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
201X

Approved by:

Committee Member 1

Committee Member 2

Committee Member 3

Committee Member 4

Committee Member 5

Committee Member 6

ABSTRACT

YOUR M. NAME: Your Title in Title Font, but not in all Caps
(Under the direction of Your Boss)

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Dedication. . .

ACKNOWLEDGEMENTS

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet,

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABD	All But Dissertation
I/O	Input/Output
IPC	Inter-Process Communication
IPI	Inter-Processor Interrupt
WSS	Working Set Size
AYO	Add Your Own in alphabetic order…

CHAPTER 1

Introduction

Network data appears widely across fields as a data structure for modeling relational information between a set of entities. In recent years, networks have become an indispensable data mining tool, as they allow for tasks such as, data visualization, clustering, and predictive modeling. Motivated by problems in fields, such as, biology, medicine, neuroscience, social science, and epidemiology, the field of network analysis has gained popularity and seeks to develop tools for understanding the associated network data. The development of these tools is rooted in a combination of techniques from statistics, computer science, physics, and mathematics. In this thesis, we will provide a comprehensive overview of networks and analysis techniques and introduce three new models/methods that will expand the types of network data that we asre able to collect and interpret.

1.1 Network Notation

1.1.1 Representing relational information

Humans frequently benefit from network applications for tasks such as, viewing relevant queries from a google search, enjoying a suggested movie on Netflix, or interacting on a social network platform. The basic building blocks of networks are nodes, representing entities in a systems, and edges, encoding connections their physical or inferred connection or similarity. Figure 1.1 shows a social network between 7 users and edges between them denoting whether they interact.

Such a network with edges simply representing whether or not a pair of nodes interact is an example of an *undirected,unweighted* network. We will use an undirected network to introduce two forms of representations for networks. For a set of N nodes, we define the $N \times N$ network adjacency

Figure 1.1: **Toy social network.** A small example of a social network, with nodes being users and edges representing connections between users. Image from `https://www.phpfox.com`

matrix, $\mathbf{A} = \{a_{ij}\}$. For a pair of nodes i and j, its corresponding adjacency matrix entry a_{ij} is defined as follows,

$$\begin{cases} a_{ij} = 1 & \textit{if node } i \textit{ and node } j \textit{ are connected} \\ a_{ij} = 0 & \textit{otherwise}. \end{cases}$$

.

Undirected networks can also be *weighted*, where the weight of an edge between a node pair encodes their extent of similarity. These edge weights are some real number and are frequently quantities such as correlation or pairwise similarity. A simple extension of $\mathbf{A}$ to an undirected, weighted network where w is the edge weight between nodes i and j computes the adjacency matrix entry a_{ij} as,

$$\begin{cases} a_{ij} = w & \textit{if node } i \textit{ and node } j \textit{ are connected} \text{ with weight } w \\ a_{ij} = 0 & \textit{otherwise}. \end{cases}$$

Alternatively, the assumption of a symmetric relationship between a pair of nodes that node i connects to node j and node j connects to node i may be unrealistic. For example, on twitter, user i can follow user j, but user j does not necessarily need to follow user i. This type of network is known as a *directed* network. While directed are frequently discussed in the network science literature, we will not introduce them here.

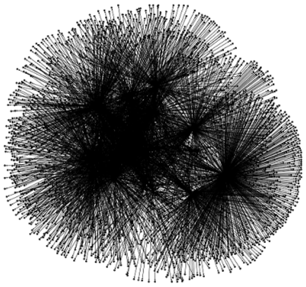

Figure 1.2: **Hairball network.** Networks are often noisy data structures and lack an immediate straight forward structural interpretation. Image from `https://cs.umd.edu`

1.1.2 Basic Analysis

Given a network, there are fundamental tasks of interest that allow for a more clear interpretation and understanding of the data. Some of these objectives include, quantifying node importance, quantifying edge density, identifying connected components ,clustering nodes, and predicting links. Networks in textbooks often look deceptively clean and well-structure. In reality, most network data is described as being a hairball. This term refers to the difficulty of discerning structure or interpreting meaning from the network based on the connectivity patterns. An example of a typical hairball is shown in figure 1.2

Such a challenging representation of the data requires breaking the network down into smaller pieces that can be further analyzed. The first most basic summary statistic is known as *degree*. Here, we will define a variety of summary statistics and quantities that can be computed on a network that give insight into the network's structure. Given the adjacency matrix for an undirected network, $\mathbf{A}$, the degree of node i, degree(i) is computed as,

$$\text{degree}(i) = \sum_j a_{ij} \tag{1.1}$$

In the case of an undirected, unweighted network, the degree of node i counts its number of neighbors, while in the undirected, weighted context, degree encodes the total edge weight incident to node i. Collectively examining the distribution of degrees for a network is known as the *degree distribution*. Understanding the degree distribution provides insight into the network type and structural organization. [Add some example maybe]. To concisely summarize this information, one

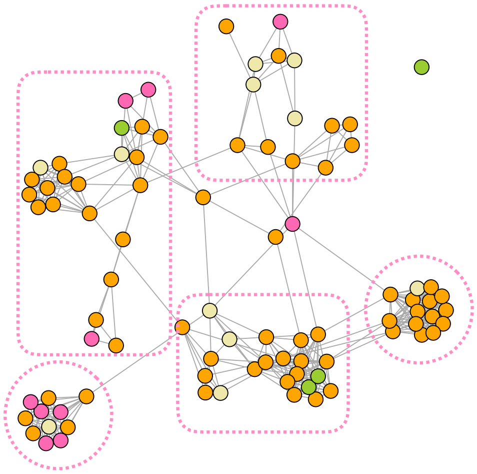

Figure 1.3: **Assortative Community Structure.** Nodes are tightly connected to each other and more sparsely connected to the rest of the network. Each community is outlined with a pink dotted line.

may consider. ... blah blah to add. Finally, clustering on a network or identifying a partition of nodes into groups or ‘communities’ based on structural network patterns is known as community detection. This is a powerful way to segment a network into smaller structures that can be further prioritized for additional analysis.

1.2 Conceptual Overview of Community Detection

A community in a network is broadly defined as a set of who share something in common in terms of their connectivity patterns in the network. One can think of a community as a clustering problem on networks, where the objective is to identify a set of nodes that are highly similar. The most basic type of community to understand is a network with assortative community structure. In this case, nodes are tightly connected to each other but more sparsely connected to the rest of the network. An example of a network with assortative community structure is shown in **??** Communities in the network are outlined with pink dotted lines.

Alternatively, networks can have a dissasortative structure where the between community edge density exceeds the within-community density. Finally, a core periphery structure can arise when there is a central core in the network that connects to the rest of the network and a set of peripheral nodes that connect to the core, but not to each other.

Community detection is a well-studied sub-domain of network science. The interested reader can refer to one of the comprehensive review articles (Lancichinetti and Fortunato, 2009; Fortunato and Hric, 2016; Shai et al., 2017)

1.3 The state of the art methods

When performing community detection on a network, the objective is to segment nodes into one of K communities. This K can be known apriori or estimated through some kind of model selection or quality function computations. There are many optimization approaches that can be used to approach network community detection. In this section, we will introduce the current state-of-the-art approaches characterized as quality function maximization, deep learning, higher order clustering, probabilistic, and spectral methods. These methods are discussed based on their ability to handle networks of non-trivial size with diverse structures.

1.3.1 Quality function maximization with modularity

For quality function optimization, one writes down a quantity to optimize that seeks to identify a partition of the network into nodes that is representative of the network structure. The most common quality function for this task is known as modularity (Newman, 2006). Intuitively, modularity defines a null model for network that doesn't have prominent organizational structure. In particular, this null model is a random graph model, known as the configuration model (Bender and Canfield, 1978). To generate an N-node network from the configuration model, one first specifies a fixed degree sequence, $D = \{k_i, k_2, \ldots, k_N\}$. From this sequence, nodes are connected with k_i stubs that will ultimately be connected together. Finally, the graph is constructed by randomly choosing pairs of the crreated stubs and joining them. Based on how this network was generated, it is easy to specify the probability that an edge exists between a pair of nodes, i and j, or $p(a_{ij} = 1$.

$$p(a_{ij} = 1) = \frac{k_i k_j}{2M}. \tag{1.2}$$

Here, k_i and k_j represent the number of edges for nodes i and j, respectively, and M is the total number of edges in the network.

Modularity was introduced in 2004 by Newman and Girvan (Newman and Girvan, 2004). We define the modularity quality function, Q as,

$$Q = \frac{1}{2M} \sum_{i,j} \left[a_{ij} - \gamma \frac{k_i k_j}{2M} \right] \delta(z_i, z_j) \tag{1.3}$$

Here, γ is a resolution parameter (Reichardt and Bornholdt, 2006) that controls the scale of community size. Large values of γ favor more small communities while smaller value enforce for fewer large communities.

In order to determine $\mathbf{z}$, the most computationally efficient approach is known as the Louvain algorithm (Blondel et al., 2008). The Louvain algorithm is an agglomerative heuristic, which initially starts with each node in its own community and in the first match merges pairs of nodes if their merge leads to an increase in modularity. Each group of nodes assembled after this first pass becomes a new node in the network and a new weighted network is created between the set of new nodes. The weight on the edges of the new network are the number of edges from the original network that go between the sets of merged nodes. This process is continues iteratively until the modularity no longer increases. The reason that this approach is so computationally tractable is because the gain in modularity, ΔQ of merging two groups of nodes can be explicitly computed in closed form.

Modularity has shown to be effective in applications from neuroscience (Meunier et al., 2009) to image segmentation (Browet et al., 2011).

1.3.2 Identifying communities with probabilistic approaches

This approach will be only briefly introduced here, as it will be explored more in depth in subsequent chapters. Probabilistic community detection methods aim to find a partition of the network through likelihood optimization. Intuitively, the goal is to study the generative process of the node edges in terms of inferred community assignments. For example, given nodes i and j, one may model $P(a_{ij} = 1)$ as $g(z_i, z_j)$, where $g(\cdot)$ is some rule based on the node-to-community assignments. Two common probabilistic community detection models are the stochastic block model (Snijders and Nowicki, 1997) and the affiliation model (Yang and Leskovec, 2012). The definition and description of these models and inference techniques are described in depth in chapter 2.

1.3.3 Deep Learning Approaches

In recent years, deep learning has begun to revolutionize many fields, including network analysis. Perozzi *et al.*, pioneered the use of deep learning in community detection with the development of DEEPWALK (Perozzi et al., 2014) to learn a latent space representation of nodes in a lower dimensional space (i.e. an emedding). Once the network is embedded in a lower dimensional space, simple clustering techniques, such as k-means (Hartigan and Wong, 1979) can be used to partition the network into communities. The approach to learn an embedding for the network is based on random walks on the network (Noh and Rieger, 2004; Gleich, 2015). A random walk on a network involves choosing a starting node and traversing the network by hopping between adjacent nodes. The DEEPWALK approach seeks to learn an embedding of the nodes that preserves the sets of nodes traversed in a random walk. To do this, the authors used Word2Vec, a tool from natural language understanding that allow for the specification of a node embedding that enable accurate prediction of a word's context, given the word (Mikolov et al., 2013). To adapt this context to networks, a random walk is treated as a sentence and nodes are treated as a word within the sentence. Moreover, the analogous task to the problem in text data to a network is to accurately assign a probability predict a set of nodes likely to be seen with the node of interest. Moreover, this problem is solved using the same optimization approach as Word2Vec

Based on the success of DEEPWALK, the method was followed up with Node2Vec in 2016 (Grover and Leskovec, 2016). While node2vec also uses the random walk framework to specify the optimization problem, they modify how the random walk is performed to enable an embedding that captures different aspects of a potential network community. For example, one may describe a community by a set of nodes located close to each other in the network with many common neighbors and connections to common neighbors. This assumption is known as network homohpily (Kossinets and Watts, 2009). Alternatively, perhaps a good definition of a community is a set of networks that have similar roles in the network. This idea is known as structural equivalence (Lorrain and White, 1971). For example, a grouping of nodes that take into account their degree, with the community assignments being highly related to node degree. To modify the random walk so that it leads to a model that gives flexibility in the nature of retrieved communities, the authors introduced a search bias term, which controls whether the random walk in performed in a breadth-first or depth-first search parameter. If on a random walk, the path is traversed in a depth-first search, favoring the exploration of a larger area of the network far from the source, the resulting community aligns with the homophily hyptohesis. A random walk performed in a breadth first manner that restricts the path to nodes neighboring the source and tends to capture nodes based on structural equivalence (i.e. a hub, or highly connected node).

1.4 Case studies in network community detection

A community approach to network analysis has shown to be fruitful in particular, in the analysis of biological and brain connectivity applications. In this section, we will describe examples of analyses where the identification of communities provided insight and understanding for a scientific problem.

1.4.1 Network analysis in computational biology

Multiple experimental modalities exist that enable the collection and analysis of biological data. Understanding protein expression, gene expression, microbiome composition, metabolomic profiles, genomic mutations, and immune profiling are just a few of examples of biological data that is studied routinely for insight into human health. With most experimental platforms producing high dimensional data, it is crucial to have good tools for interpretation, visualization, and prediction. Machine learning techniques in computational biology have revolutionized prediction in healthcare and medicine. Here, we outline particular examples of how community detection lead to important biological understanding and predictive ability.

Immunological profiling to establish a pregnancy immune clock A study lead by Aghaeepour *et al.*, demonstrated that there is a typical timing of immunological events in a healthy, term, human pregnancy (Aghaeepour et al., 2017). Immunological profiling was performed on a training cohort of 18 women, using a technology called mass cytometry (Bendall et al., 2012) was used to quantify various features of the immune system, such as, cell type abundances, signaling activity. From this set of measured immune features, a correlation network from the training cohort to identify which immune features were potentially related or working together. Simultaneously, a regression model was training to identify immune features associated with increased gestational age. When communities were identified in the network of immune features, there were two important observations. First, immune features of the same type (i.e. cell signaling vs. cell frequency) were aligned with community labels. Second, sets of features associated with a particular gestation age often fell in the same community, indicating their synchronous activity during the pregnancy. Finally, after identifying influential nodes in their ability to predict stage in pregnancy, according to the regression model, the communities of these nodes were more closely examined to uncover further insight into the immunological mechanisms occuring throughout the pregnancy time course.

Uncovering differences in microbiome community structure in patients with inflammatory bowel disease

The microbiome refers to the collection of bacterial species that populate an organism's gut. Microbiome analysis has recently gained attention, as its biological implications are large for health and disease. A 2017 review article presented the idea that the development of network analysis approaches for microbiome data is under explored and has great potential for advancing biological understanding and interpretation of these data (Layeghifard et al., 2017). A network in this context is typically constructed based on some notion of co-occurence or correlation between microbial species, profiled across samples A recent example where community detection played a key role in the biological understanding was introduced in 2017 and assessed the interplay between microbial co-occurence structural organization patterns between patients with and without inflammatory bowel disease (Baldassano and Bassett, 2016). Communities were identified in the healthy and diseased networks, using classic modularity maximization (**?**). After identifying a community structure for each network, the similarity of these partitions was quantified with the Rand index (Traud et al., 2011), which showed to be statistically significant under a permutation test. This observation allowed the authors to understand that the core structure from a healthy microbiome was conserved even in diseased patients, but allowed for more careful probing of the subtle differences. First, the functional roles of the members of each community were interrogated. Some interesting co-occurence relationships within communities were identified, such as the loss of strong clustering, or association propensity between pro and anti-inflammatory species within the diseased networks. This interplay between pro and anti inflammatory species is thought to play a pivotal role in the maintenance of a healthy gut microbiome. [ADD ABOUT node roles]

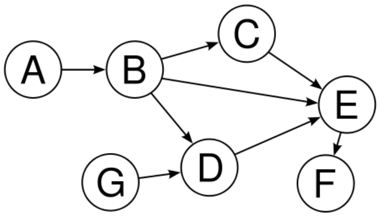

Figure 1.4: **Directed Acyclic Graph.** A directed acyclic graph (DAG) is formed based on dependency between random variable and allows for a fully factorized probability distribution.

1.5 Network analysis in neuroscience

1.6 Network analysis software

1.7 Open problems in community detection

1.8 Probabilistic graphical models for statistical inference

Probabilistic network models are one approach to community detection that seek to model edge existence based on the node-to-community assignments. In doing so, the objective is to learn the node-to-community assignments that make the structure of the observed network the most likely. In this section, we will define some useful notation and concepts To fit a probabilistic network model to data, we will define some useful notation and concepts that help simplify writing down and interpreting the likelihood.

Probabilistic graphical models enable efficient specification and manipulation of large probability distributions through semantic structures. Given a set of random variables, $\{A, B, C, D, E, F\}$, we seek to compute the joint distribution, $P(A, B, C, D, E, F)$. This joint distribution can be expressed with a directed acyclic graph (DAG), whose structure encodes dependencies between random variables. The DAG allows for the representation of the joint distribution in a factorized way, which is computationally useful. A DAG between the set of random variables, $\{A, B, C, D, E, F\}$ is shown in 1.4.

To translate a DAG between a set of N random variables, $\mathbf{X} = \mathbf{X} = \{X_1, X_2, \ldots, X_N\}$ to its joint distribution, we rely on the Factorization theorem, which specifies that a DAG factors according to its parent/child relationships with,

$$P(\mathbf{X}) = \prod_{i=1:N} P(X_i \mid \mathbf{X}_{\pi_i}). \tag{1.4}$$

Here, π_i denotes the set of parents for node i. Using this information, we can write down the joint distribution for figure 1.4 as,

$$P(A, B, C, D, E, F) = P(A)P(B \mid A)P(C \mid B)P(D \mid B, G)P(E \mid D, B, C)P(F \mid E). \tag{1.5}$$

This introduced idea will help in subsequent sections to expresses a model graphically, write down the model likelihood, and use the likelihood to optimize for the most appropriate model parameters.

CHAPTER 2

Probabilistic community detection models and inference techniques

In this section, we will present two probabilistic models for community structure, the stochastic block model and the affiliation model.

2.1 Stochastic block model

2.1.1 Most general stochastic block model

For an undirected, unweighted network $\mathcal{G}$ with adjacency matrix, $\mathbf{A}$, we seek to partition each of the N nodes into one of K communities. We denote the the node-to-community assignments as $\mathbf{z}$, with z_i specifying the community assignment of node i. Here, $\mathbf{z}$ is a latent variable, with each entry taking on 1 of K states, or one of K community assignments. Figure 2.1 shows the dependency relationship between the node-to-community assignments. Here, the node-to-community assignments are treated as a latent variables because we seek to identify the $\mathbf{z}$ that makes the observed adjacency matrix, $\mathbf{A}$ the most likely. The crucial assumption of the stochastic block model is that nodes within a community are connected to nodes within their community and to other communities in a characteristic way. To this end, the model fitting procedure requires learning a set of within and between community connection probabilities. Under this approach, edges are treated as independent and identically distributed and deciding whether or node an edge exists between a pair of nodes is the learned connection probability between the communities to which each of the nodes belong.

Using the factorization rules described in section 1.8, we can specify the complete data log likelihood between $\mathbf{z}$ and $\mathbf{A}$ as,

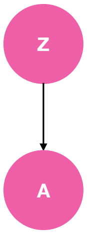

Figure 2.1: **SBM Graphical Model.** A graphical model is used to model the dependency between the node-to-community assignments, $\mathbf{z}$ and the observed network adjacency matrix, $\mathbf{A}$.

$$\log P(\mathbf{z}, \mathbf{A}) = \log(P(\mathbf{A} \mid \mathbf{z})) + \log(P(\mathbf{z})) \tag{2.1}$$

To further specify these communities, we will define additional notation. First, let $\mathbf{\Pi}_{K \times K} = \{\pi_{ij}\}$ be the matrix that specifies the within and between community edge probabilities. Using this information, we can model the probability of an edge existing between nodes i and j as,

$$P(A_{ij} = 1) \sim \text{Bernoulli}(\Pi_{z_i, z_j}) \tag{2.2}$$

We let $Z_i = \{Z_{i1}, Z_{i2}, \ldots Z_{ik}\}$ be a collection of binary indicators where Z_{ik} is 1 i belongs to community k and 0 otherwise, We also let α_k be the probability that a node belongs to community k. With all of this information, we can write down each term of the complete data likelihood.

First,

$$\log(P(\mathbf{Z})) = \sum_i \sum_k Z_{ik} \log(\alpha_k). \tag{2.3}$$

Next,

$$\log(P(\mathbf{A} \mid \mathbf{Z})) = \sum_{i \neq j} \sum_{k < l} Z_{ik} Z_{il} [a_{ij} \log(\Pi_{kl}) + (1 - a_{ij}) \log(1 - \Pi_{kl})] \tag{2.4}$$

Optimizing the parameters of this incomplete data log likelihood requires computing the posterior $P(\mathbf{z} \mid \mathbf{A})$ but as shown by (Daudin et al., 2008) is intractable. To address this issue, the posterior can be recast using a factorized approximation. This is accomplished by optimizing a lower bound of $\mathcal{L}(\mathbf{A})$. We let $\mathcal{R}_A$ be an approximation of the posterior, $P(\mathbf{z} \mid \mathbf{A})$. To optimize the lower bound of

$\log \mathcal{A}$, we seek the $\mathcal{R}_A$ that is as close as possible to $P(\mathbf{z} \mid \mathbf{A})$. In other words, we define the lower bound of $\mathcal{L}(\mathbf{A})$ as $\mathcal{T}(\mathcal{R}_A)$, with,

$$\mathcal{T}(\mathcal{R}_A) = \log \mathcal{L}(\mathbf{A}) - \mathrm{KL}[\mathcal{R}_A(\mathbf{z}), \mathbf{P}(\mathbf{z} \mid \mathbf{A})]. \tag{2.5}$$

Here KL denoted the Kullback-Leibler divergence (KL divergence) and the best approximation will be the value that makes the KL divergence the smallest. Jaakkola *et al.*, present a mean field approximation for the posterior distribution (Jaakkola, 2001) as,

$$\mathcal{R}_A(\mathbf{z}) = \prod_i h(Z_i; \boldsymbol{\tau}_i). \tag{2.6}$$

Here $\boldsymbol{\tau} = (\tau_{i1}, \ldots, \tau_{iK})$ and τ_{ik} is the approximation that node i belongs to community k, or $P(Z_{ik} = 1 \mid \mathbf{A})$. Furthermore, $h(\cdot; \boldsymbol{\tau}_i)$ denotes the multinomial distribution with parameter $\boldsymbol{\tau}$.

Daudin (Daudin et al., 2008) *et al.*, show that the optimal estimate for τ_{ik} denoted $\hat{\tau}_{ik}$ satisfies

$$\hat{\tau}_{ik} \propto \alpha_k \prod_{j \neq i} \prod_{l} [\theta_{z_i, z_j}^{a_{ij}} (1 - \theta_{z_i, z_j})^{1 - a_{ij}}]^{\hat{\tau}_{ik}}. \tag{2.7}$$

Here, α_k notes the probability that a node belongs to community k. Furthermore, after computing the set of variational parameters, the updates for $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ that maximize $\mathcal{T}(\mathcal{R}_A)$ are also shown by Daudin *et al.,* (Daudin et al., 2008) to be,

$$\hat{\alpha}_k = \frac{1}{n} \sum_i \hat{\tau}_{ik} \qquad \theta_{ql} = \sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jl} a_{ij} / \sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jl} \tag{2.8}$$

We have presented this variational approach for performing SBM parameter inference and likelihood optimization because this approach was appropriate for the work presented in this thesis. Variational inference is just one approach that can be applied to learn model parameters and was but a study by Zhang *et al.* (Zhang et al., 2012) also show that belief propgation is very effective for this task (Murphy et al., 1999). Briefly, belief propagation is a message passing algorithm for parameter inference in probabilistic graphical models. Given that parameter learning offer requires computing marginal distributions for a set of variables with a very large number of possible configurations, belief propagation uses the graphical model to reduce the complexity of the problem. Using the belief propagation to infer latent node-to-community assignments and update the model parameters was shown to perform superperior to the variational appromixation

This formulation of the problem and parameter optimization procedure works well and converges quickly for networks that have assortative community structures and a homogenous degree distribution. We will now explore how this classic formulation of the SBM can be modified to enable a broader application for a variety of networks.

2.1.2 Variants to the Classic Stochastic Block Model

The introduced stochastic block model is the most vanilla version in that it makes the assumption that the network is unweighted, each node is assigned to only one community. The introduced model also does not account for issues that may arise from degree heterogeneity (i.e. a large disparity in node degree in sets of nodes). Here, we will briefly discuss the approaches that adapt the stochastic block model to handle these issues and assumptions.

Edge Weights

The majority of the stochastic block model literature considers unweighted networks simply because describing a probabilistic model to handle both edge existence and edge weight is a challenging task. In the classic stochastic block model, we are simply modeling whether an edge exists based on the inferred community memberships of the edge stubs. Since edge weights can come in a variety of forms (real-valued, count, etc.), it is difficult to immediately decide what distribution the edge weights should follow. In the past few years, this issue has been tackled in two papers (Aicher et al., 2014; Peixoto, 2018).

First, Aicher *et al.* developed a model and associated inference technique, for the weighted stochastic block model. Here, edge weights can be modeled by any exponential family distribution. The authors use a mixing parameter that allows for the control of the use of edge existence versus edge weights when learning node-to-community assignments. This method requires having an estimate of the number of communities, K, but the paper provides an approach to use Bayes' factors between two competing values of K to determine which model is a better fit.The inference for fitting this model is performed through a variational bayes approach (Attias, 2000).

To avoid having intuition about K, Peixoto (Peixoto, 2018) developed a non parametric bayesian approaches that is capable of inferring K with no prior knowledge. The assumption of the model is also slightly different and assumes a hierarchical structure between communities. The inference is achieved through MCMC sampling.

Degree Heterogeneity

Based on the variety of network structures and types, the assumption that the classic stochastic block model is an appropriate model for the data is often invalid. That is, for some networks, the fitted model may not actually be a good fit for the data. Work by Karrer *et al.*, introduced a simple extension to the classic stochastic block model, known as the degree corrected stochastic block model, that is informed by degree distribution as a proxy for the network structure. In networks where there is a high disparity between node degree (i.e. many high degree nodes and many low degree nodes), stochastic block models inference tends to partition the nodes intro communities of high degree and low degree nodes. The approach for adapting the SBM to this setting is to learn a $K \times K$ matrix, $\boldsymbol{\theta}$, describing the number of edges between each pair of communities. these counts are modeled as poisson random variables. The likelihood of the observed network under this poisson assumption takes into account node degrees.

The restriction of single community membership

As it is often observed in social networks, the assumption that every node belongs to only a single community is restrictive. To address this issue, approaches have been developed to allow nodes to participate in a mixture of communities (Airoldi et al., 2008) or to overlapping groups (Latouche et al., 2011). Airoldi *et al.*, pioneered the development of the mixed membership stochastic block model (Airoldi et al., 2008), where instead of modeling a node's membership in each community in a binary manner, the authors allow a node to belong to multiple communities. The generative process for this approach for modeling the existence of an edge between nodes p and q in a network with K possible communities and $\boldsymbol{\theta}$ representing the between community connection probabilities.

- For each node p, draw a mixed membership vector $\pi_p \sim$ Dirchelet($\boldsymbol{\alpha}$)
- Then for each pair of nodes (p, q), draw $\mathbf{z}_{p \to q} \sim$ Multinomial(π_p), $\mathbf{z}_{q \to p} \sim$ Multinomial(π_q)
- Sample the edge between p and q as, A_{pq}, where $A_{pq} \sim$ Bernoulli($\mathbf{z}_{q \to p}^T \boldsymbol{\theta} \mathbf{z}_{q \to p}$)

Following the development of the mixed membership stochastic block model, Latocuhe *et al.* (Latouche et al., 2011) addressed an important limitation of (Airoldi et al., 2008). Since the probability of an edge between a pair of nodes p and q depends on a single draw of $\mathbf{z}_{p\rightarrow q}$ and $\mathbf{z}_{q\rightarrow p}$, the class memberships of nodes p and q towards other nodes in the network are ignored. Moreover, this model adapts the mixed membership stochastic block model to incorporate more structures of the network.

2.2 Affiliation model and inference

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer

CHAPTER 3

Community Detection in multilayer networks

CHAPTER 4

A multilayer stochastic block model

4.1 Fitting a common stochastic block model to all network layers

4.2 Strata multilayer stochastic block model

4.3 Parameter learning

4.4 A clustering-based fitting approach

4.5 Synthetic examples

4.6 Detectability limits

4.7 Human microbiome project example

4.8 Comparison to reducibility

CHAPTER 5

Network compression

CHAPTER 6

Network compression for community detection with super nodes

6.1 Super pixel pre-processing of images

6.2 Super node pre-processing for networks

6.3 2-Core decomposition approach for selecting seeds as community centers

6.4 Creating a super node network representaion

6.5 Social network data examples

6.6 Benefits of a compressed representation: run time, variability, neighborhood smoothing

CHAPTER 7

Attributed networks and community detection

7.1 Examples of attributed networks

7.2 Models and inference for attributed networks

7.3 Alignment of attributes with communities

CHAPTER 8

An attributed stochastic block model

8.1 Approaches to an attributed stochastic block model

8.2 A model of conditional independence between attributes and connectivity

8.3 Learning the model parameters

8.4 Example on a synthetic attributed network

8.5 Detectability limits in attributed networks

8.6 Case studies for attributed networks

8.7 Attributed SBM in link prediction

8.8 Attributed SBM in collaborative filtering

CHAPTER 9

Software

9.1 sMLSBM

9.2 Super node representations for a network

9.3 Attributed stochastic block model

CHAPTER 10

Conclusion and future work

BIBLIOGRAPHY

Aghaeepour, N., Ganio, E. A., Mcilwain, D., Tsai, A. S., Tingle, M., Van Gassen, S., Gaudilliere, D. K., Baca, Q., McNeil, L., Okada, R., et al. (2017). An immune clock of human pregnancy. *Science immunology*, 2(15):eaan2946.

Aicher, C., Jacobs, A. Z., and Clauset, A. (2014). Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248.

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

Attias, H. (2000). A variational baysian framework for graphical models. In *Advances in neural information processing systems*, pages 209–215.

Baldassano, S. N. and Bassett, D. S. (2016). Topological distortion and reorganized modular structure of gut microbial co-occurrence networks in inflammatory bowel disease. *Scientific reports*, 6:26087.

Bendall, S. C., Nolan, G. P., Roederer, M., and Chattopadhyay, P. K. (2012). A deep profiler's guide to cytometry. *Trends in immunology*, 33(7):323–332.

Bender, E. A. and Canfield, E. R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307.

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*, 2008(10):P10008.

Browet, A., Absil, P.-A., and Van Dooren, P. (2011). Community detection for hierarchical image segmentation. In *IWCIA*, volume 11, pages 358–371. Springer.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and computing*, 18(2):173–183.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

Gleich, D. F. (2015). Pagerank beyond the web. *SIAM Review*, 57(3):321–363.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Jaakkola, T. (2001). 10 tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, page 129.

Kossinets, G. and Watts, D. J. (2009). Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117.

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

Layeghifard, M., Hwang, D. M., and Guttman, D. S. (2017). Disentangling interactions in the microbiome: a network perspective. *Trends in microbiology*, 25(3):217–228.

Lorrain, F. and White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1):49–80.

Meunier, D., Lambiotte, R., Fornito, A., Ersche, K. D., and Bullmore, E. T. (2009). Hierarchical modularity in human brain functional networks. *Frontiers in neuroinformatics*, 3.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Noh, J. D. and Rieger, H. (2004). Random walks on complex networks. *Physical review letters*, 92(11):118701.

Peixoto, T. P. (2018). Nonparametric weighted stochastic block models. *Physical Review E*, 97(1):012306.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM.

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1):016110.

Shai, S., Stanley, N., Granell, C., Taylor, D., and Mucha, P. J. (2017). Case studies in network community detection. *arXiv preprint arXiv:1705.02305*.

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100.

Traud, A. L., Kelsic, E. D., Mucha, P. J., and Porter, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543.

Yang, J. and Leskovec, J. (2012). Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1170–1175. IEEE.

Zhang, P., Krzakala, F., Reichardt, J., and Zdeborová, L. (2012). Comparative study for inference of hidden classes in stochastic block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(12):P12021.