COMMUNITY DETECTION IN MULTIMODAL NETWORKS

Mark (Tianshe) He

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 Statistics and Operations Research.

Chapel Hill
2021

Approved by:

Shankar Bhamidi

Nikhil Kaza

J.S. Marron

Andrew Nobel

Vladas Pipiras

ABSTRACT

Mark (Tianshe) He: Community Detection in Multimodal Networks
(Under the direction of Shankar Bhamidi and Andrew Nobel)

Community detection on networks is a basic, yet powerful and ever-expanding set of methodologies that is useful in a variety of settings. This dissertation discusses a range of different community detection on networks with multiple and non-standard modalities. A major focus of analysis is on the study of networks spanning several layers, which represent relationships such as interactions over time, different facets of high-dimensional data. These networks may be represented by several different ways; namely the few-layer (i.e. longitudinal) case as well as the many-layer (time-series cases). In the first case, we develop a novel application of variational expectation maximization as an example of the *top-down* mode of simultaneous community detection and parameter estimation. In the second case, we use a *bottom-up* strategy of iterative nodal discovery for these longer time-series, abetted with the assumption of their structural properties. In addition, we explore significantly self-looping networks, whose features are inseparable from the inherent construction of spatial networks whose weights are reflective of distance information. These types of networks are used to model and demarcate geographical regions. We also describe some theoretical properties and applications of a method for finding communities in bipartite networks that are weighted by correlations between samples. We discuss different strategies for community detection in each of these different types of networks, as well as their implications for the broader contributions to the literature. In addition to the methodologies, we also highlight the types of data wherein these "non-standard" network structures arise and how they are fitting for the applications of the proposed methodologies: particularly spatial networks and multilayer networks. We apply the *top-down* and *bottom-up* community detection algorithms to data in the domains of demography, human mobility, genomics, climate science, psychiatry, politics, and neuroimaging. The expansiveness and diversity of these data speak to the flexibility and ubiquity of our proposed methods to all forms of relational data.

To my family

ACKNOWLEDGEMENTS

I thank my advisors as well as my committee for the obvious things of mentorship, advising, training, identifying important outstanding problems in the literature, etc. I want to thank Shankar and Andrew especially for their unwavering support and for having the confidence in me even in the face of obstacles. I thank Steve for his advice and wisdom imparted in the excellently designed statistical consulting class. I thank Nikhil for imparting valuable knowledge not only his domain expertise in geography and urban planning but also in statistical computation. I thank Vladas for his continued support not only in research development but also in personal growth and wellness.

I thank Professor Rose Xavier and the Xavier lab for providing PNC data and mentorship and for collaboration in clinical applications to analyzing schizophrenia pathology. I also thank Professor Jason Xu for collaboration and guidance on the work on multilayer block models. I also thank Professor Galen Reeves for helpful advice in contextualizing this work to the literature. I thank the helpful comments of Professor Richard Smith for the climate application of the Bimodules Clustering algorithm. I also want to thank Professor Don Hedeker for giving very helpful advice about research and career. I wish to acknowledge the generous funding offered by the NDSEG fellowship that supported me from 2017-2021.

I also wanted to especially thank the amazingly supportive staff in the STOR department: Alison Kieber, Christine Keat, Sam Radel, Shayna Hill, and Ayla Ocasio. I could not have done this without them.

The data for the project described in Chapter 2 was funded by the Rockefeller University Heilbrunn Family Center for Research Nursing (RX, 2019) through the generosity of the Heilbrunn Family. The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. Philadelphia Neurodevelopment Cohort (PNC) clinical phenotype data used for the analyses described in this manuscript were obtained from dbGaP at `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap` through

dbGaP accession `phs000607.v3.p2.` Support for the collection of the data for Philadelphia Neurodevelopment Cohort (PNC) was provided by grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson. Subjects were recruited and genotyped through the Center for Applied Genomics (CAG) at The Children's Hospital in Philadelphia (CHOP). Phenotypic data collection occurred at the CAG/CHOP and at the Brain Behavior Laboratory, University of Pennsylvania.

I thank the mentorship and advice of students in the department, namely Eric Friedlander, John Palowitch, Dylan Glotzer, Zhengling Qi, Hongsheng Liu, Iain Carmichael, and Jon Williams. I especially want to thank Eric and John for directly helping with some of the ideas presented in this dissertation. I want to thank the others in my graduate cohort for their friendship and support, especially Michael Conroy, Adam Waterbury, Jack Prothero, Aman Barot, Miheer Dewaskar, and Samopriya Basu. I also want to thank the friends I have formed communities with in the Research Triangle. I also want to thank and express my continued gratitude for the community institutions therein.

PREFACE

Our data becomes us.

– Prof. Alexandra Chassanoff

We are always, however, brought back to a dissymmetrical necessity to cross from the smooth to the striated, and from the striated to the smooth. If it is true that itinerant geometry and the nomadic number of smooth spaces are a constant inspiration to royal science and striated space, conversely, the metrics of striated spaces (metrons) is indispensable for the translation of the strange data of a smooth multiplicity. Translating is not a simple act: it is not enough to substitute the space traversed for the movement; a series of rich and complex operations is necessary. Neither is translating a secondary act. It is an operation that undoubtedly consists in subjugating, overcoding, metricizing smooth space, in neutralizing it, but also in giving it a milieu of propagation, extension, refraction, renewal, and impulse without which it would perhaps die of its own accord: like a mask without which it could neither breathe nor find a general form of expression. Major science has a perpetual need for the inspiration of the minor; but the minor would be nothing if it did not confront and conform to the highest scientific requirements.

–Deleuze and Guattari, *A Thousand Plateaus*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BSP	Bimodules search procedure
CCME	Continuous configuration model extraction
CCME-SL	Continuous configuration model extraction - self looping
ELBO	Evidence lower bound
SBM	Stochastic blockmodel
SBANM	Stochastic block (with) ambient noise model
VI	Variational inference
VEM	Variational expectation maximization
α	Significance level
α_q	Membership probability (Chapter 2)
$\boldsymbol{\mu}_q$	Mean of community indexed at q
$\boldsymbol{\mu}_{AN}$	Mean of *ambient noise* distribution
$\boldsymbol{\Sigma}_q$	Variance of community indexed at q
$\boldsymbol{\Sigma}_{AN}$	Variance of *ambient noise* distribution
ρ_q	Correlation indexed at q
ρ	Self-looping tendency

CHAPTER 1
Introduction

As more information becomes available in the age of "big data", methods to gather, process, and derive meaning from these data have become more relevant and urgent in the face of escalating global problems. Relational data have become more common in the advent of sophisticated data gathering mechanisms and more nuanced conceptions of dependency. Statistical network analysis has become a major field of research and is a useful mode of pattern discovery. Networks representing social interactions, genes, and ecological webs often model members or agents as nodes (vertices) and their interaction as edges. General references on statistical modeling of random graphs include the recent books by Newman et al. (Newman, 2018a) and user manual by Fortunato (Fortunato and Hric, 2016).

Alongside the study of networks, the field of *community detection* on networks has grown considerably in recent times, with a host of methodologies from many different fields including computer science, physics, and statistics. Broadly put, community detection is an approach used to divide a set of nodes in a given relational structure into clusters whose members are strongly connected. Many techniques have been proposed for *unweighted* or binary data including modularity optimization (Girvan and Newman, 2002; Clauset et al., 2005), stochastic block models (Holland et al., 1983; Nowicki and Snijders, 2001; Peixoto, 2018; Yan et al., 2014), and extraction methods (Zhao et al., 2012; Lancichinetti et al., 2011). The three general modes of clustering for networks in the literature can gnerally be categorized under (1) optimization, (2) statistical inference, and (3) null model dynamics.

This dissertation is primarily about community detection in different types of weighted networks. Ambitiously, the goal of this dissertation is to expand on the frontier of clustering in modern network-structured data using a constellation of different novel methodologies. Though canonical methods with associated theoretical postulations are largely focused on simple networks, much data in recent years possess more complex forms. These models directly map to the variegation of ways

that data interact and relate with one another, and are particularly synchronous with the rise of availability in more different types of data, with even more complex configurations of communal structures amongst them.

We present methods of community detection on networks that represent several facets of real-world data. Specifically, we focus on the practical problem of discovering interconnected clusters amongst networks describing processes over differing time periods or modalities, as well as in networks whoses vertices represent relationships that are "irregular" compared to typical literature in network science (i.e. spatial points). Our contributions to the field of community detection serve several different practical functions. These methods fall in the *optimization, inference*, and *null model* categories of the previously mentioned general domains of approaches. The first major part of this dissertation in Chapter 2 uses variational inference to estimate the memberships and parameters of a multilayer weighted stochastic block structure with global ambient noise. The second major part of this dissertation primarily engages in *significance testing* methodology in Chapters 3-5, which iteratively uses the measures of significance between observations (nodes) and sets and rejects the connected nodes until the memberships of the candidate set reaches stability. I use iterative significance testing in several different ways that hone in on the advantage of its flexibility. I apply this approach for both the self-loop-accounting community detection methodology in Chapter 3 and the intertemporal community detection method described in Chapter 4. Iterative testing posits a candidate set at a given iteration, then constructing a test statistic comparing the weights between the sets and nodes that potentially border the set. In Chapter 5 I describe another method also based on iterative testing for clustering bipartite datasets based on correlations between the two variable sets.

Future work will mostly comprise methodological extensions from all of the chapters. I propose some extensions for the `SBANM` in the form of a naive prediction method by means of distance-based classification to ascertain a cross-validated prediction error. In the clinical analysis of psychosis in PNC data, the method is clinically impactful because it is response-free (i.e. does not require a dependent variable) and hence allows for early detection of psychosis . We compare and conflate (parts of) this approach with other methods and aim to culminate in broader, more general models and methodologies for multimodal, high-dimensional relational data. Another part of the future work comprises some investigations into the spatial data that were used in Chapters 3, 4, and 5.

1.1 Background

I present an overview of related work in this section, starting from the general themes that are prevalent throughout the literature on networks through various perspectives from differing disciplines. In community detection, many methods utilize quality functions such as modularity (Girvan and Newman, 2002). One of the central techniques in existing methods is modularity optimization wherein scores of community assignments are optimized. However, a number of problems plague modularity, most prominently of which inherent bias in estimtion (Bickel and Chen., 2009). As such, in this dissertation I focus mostly on alternatives to modularity in the form of *null model dynamics* and *statistical inference* modes as mentioned in the last section.

One major class of methods presented below in Chapters 3-5 is based on *null models* for networks. Typically, in these schemes one constructs a network model that preserves some aspects of the observed network (in the context of unweighted networks). The preserved characteristic is often attributes such as the degree distribution of the network (see Section 1.1.2 for details). A *scrambled network*under the *configuration model* (for example) with preserved degree distributions creates a network with no inherent clustering tendency. The observed network is then compared with this null model to extract subsets which seem more densely connected within the subset as compared to the null model. More details on null models are further described in Section 1.1.2.

Another class of methods explicitly use statistical inference for fitting empirical data to networks (either Bayesian (Peixoto, 2013, 2014, 2017) or frequentist (Newman and Leicht, 2007; Karrer and Newman, 2011; Bickel and Chen., 2009) models). We use *statistical inference* for the other major class of variational inference for stochastic blockstructures presented in Chapter 2 of this dissertation.

1.1.1 Multimodal and Multilayer Networks

The general subdomain of multilayer network analysis is explored in this dissertation in Chapter 2. In many analyses of weighted networks, relationships are assumed to be of the same type such as 'friendship'. However in modern relational data-types, we often have information regarding relationships of multiple types among members. For example, nodes represented by users in a social network such as twitter can have edges that represent 'likes', 'follows', and 'mentions'. In biological

data, particles express different aspects of interactions such as physical interactions between proteins or mitochondria, or co-expressions amongst genes. These questions are especially pertinent in psychiatric data, wherein the distinct symptomologies are clearly demarcated, and instead rely on a whole constellation of differing interacting psychopathologies properly diagnose certain conditions.

While static, or *unimodal* (single-graph) approaches have been developed decades ago first in the realm of the social sciences, and later many such theoretical models in mathematics (Bollobás, 1980; Bender and Canfield, 1978) and physics (Girvan and Newman, 2002), the literature concerning weighted, dynamic models is much more recent and require even more sophisticated methods. Modeling time-dynamic and multimodal networks is an emerging field of interest and its many broad and flexible domains of application can be described by Holme et al. (Holme, 2015) , who outline many scenarios for such model formulations. Of particular interest for clinical and experimental settings are *time-window graphs*, which describe time aspects of network evolution, and *difference graphs*, which describe differential settings or test conditions of interacting systems.

Included in the discussion of *multimodality* is Joint analysis of high-dimensional data. Typically known as multi-view (or multi-modal) analysis, this subject has received considerable attention in the literature,(Lahat et al., 2015; Meng et al., 2016; Tini et al., 2019; Pucher et al., 2019; McCabe et al., 2019) Driven by the ongoing development and application of moderate and high-throughput measurement technologies in fields such as genomics, neuroscience, ecology, and atmospheric science, researchers are often faced with the task of analyzing and comparing two or more data sets derived from a common set of samples. In most cases, different technologies measure different features, and capture different information about these samples. While data may be analyzed separately, additionally important insights can be gained from their integrated analysis. These groups are loosely termed *bimodules* (Wu et al., 2009), (Patel et al., 2010), and (Pan et al., 2019).

1.1.2 Null Models

A null model in the context of community detection is a random network model which preserves some aspects of an observed network but without any explicit community structure. Null models as a strategy for community detection were initially proposed by(Maslov and Sneppen, 2002). A basic null model for networks is known as the Erdos Renyi model, which supposes a uniform, global rate of connectivity across vertices. Another common null model used in the context of unweighted

networks is the configuration model (described in a later section). Once a null model for the network is established, communities based on the null model are groups of nodes that deviate from the baseline by being more connected to each other than expected under the null model.

Various functionals are used to measure the deviation of a set or a partition of the entire node set from the null model. The most popular among these functionals is modularity score (Girvan and Newman, 2002; Clauset et al., 2005). One can then try to optimize such scores to find the best partitions. Fosdick et al. introduced a framework for configuration models that accounts for self loops and used a modularity optimization approach for community detection as an application, but did not focus on weighted networks whose self loops account for the majority of its weights (Fosdick et al., 2018). Another null-model based approach is to assess the statistical significance of the deviation of subsets from what one would expect under the null, correcting these estimates for false-discovery rates, and then extracting communities that appear to be more significantly connected than under the null model. Several approaches have implicitly utilized the notion of deviation against the null partitions, such as likelihood ratios (Yan et al., 2014) or Bayes factors (Peixoto, 2018), but the class of methods as approached by more recent authors (Palowitch et al., 2018; Wilson et al., 2014) directly assess the significance against null models like the configuration model or permutation models (Dewaskar et al., 2020).

1.1.3 Self-Loops and Spatial Networks

Though some existing network models allow self loops, we propose a method in Chapter 3 that explicitly accounts for their effects and integrates them in a iterative testing framework. Most existing methods that allow for self loops presume that self-loops are somehow *similar* in characteristic to the other edges and do not properly account for when self loops are large, as in the case of the algorithm introduced by Palowitch et al. (Palowitch et al., 2018). Previous work using tree-based methods do mention self-loops, but few focus explicitly on strongly self looping networks, where self loops account for over half of the weights. Some authors introduced methods that uses modularity maximization rescaled by the size of the self loop (Xiang et al., 2015; Cafieri et al., 2010) . Peixoto's research on Bayesian stochastic blockmodels allow for self loops (Peixoto, 2017).

Self-loops are a unique characteristic of, and perhaps inherent to, networks that represent spatial relationships. Tobler's first law of geography states that "everything is related to everything else, but near things are more related than distant things." (Tobler, 1970). In a network setting where spatial distances are embedded within edges, zero-distance relationships (self loops) naturally induce the largest weights. Substantial work on spatial network analysis of human mobility data has been done in recent years. Barthelemy et al. (Barthélemy, 2014) conducted a general survey of spatial network models and dynamic processes on these models. Batty et al. provide a concise description of network methods specialized to the understanding of cities (Batty, 2013).

Some popular network models for urban and spatial flows are known as gravitation and radiation models are described in existing work (Ren et al., 2014; Simini et al., 2012; Sarzynska et al., 2015). Some studies use existing community detection techniques on novel geographical datasets to gain insight on how inferred communities are similar to or different from existing points of interest and how they change over time (Huang et al., 2018; Du et al., 2017). These methods on commuting behavior were used to partition synthetic cities in some work (Fujishima et al., 2019), and also used to understand the polycentric spatial patterns within cities in others (Zhong et al., 2014). For a comprehensive bibliography of models utilizing network techniques related to human mobility flow and their associated data, see the work of Pappalardo et al (Pappalardo et al., 2019).

We explore the ways in which communities may arise out of spatial networks, and in what ways may the *bottom-up* approach of iterative significance testing be advantageous for these data. We describe the novel contributions of our work on these methods in the following section.

1.1.4 Finding Clusters via Significance Testing

Much of the work described in Chapters 3, 4, and 5 of this dissertation will focus on significance testing-based community detection in networks. The general principle of the methodology follows the construction of a null model, then a procedure of setting up many hypothesis tests (usually in some pre-specified sequence of steps), then iteratively rejecting the observations whose relationships to the active set do not satisfy a significance criterion. This principle was first developed by Wilson, Nobel, and Bhamidi on a 2014 study of community detection in binary networks.

Wilson et al. use significance testing to find communities in a binary network under the assumptions of the configuration model (Wilson et al., 2014). The configuration model, as mentioned

earlier, is used as a null model for underlying randomness for a binary network wherein the degree structure, representative of the influence or "power" structure of each node. The testing procedure takes an active set B (at iteration step t) and conducts a test of significance between B with all other nodes u in the network assuming that the configuration model is approximated by a binomial distribution with parameters based on the ratio of the sums of connected edges between the nodes amongst the active set and the total sum of degrees in B. This method is known as Extraction of Statistically Significant Communities (ESSC).

More methods develop this basic premise underlying clustering methodology in different directions. Palowitch et al. as mentioned earlier, use significance-testing to find significantly connected communities in weighted networks (Palowitch et al., 2018). The premise of conducting significance tests between an active set (at a given iteration) is similar the same as that in ESSC, but the important distinction is that the method is tailored for weighted graphs. As such, Palowitch and authors have constructed a weighted (or continuous) configuration model that preserves both expected strengths (sum of weights) as well as degrees. This method is known as Continuous Configuration Model Extraction (CCME).

Bodwin et al. used significance testing to find clusters of significantly differential correlations in datasets with differential conditions (Bodwin et al., 2015). Rather than taking a network-valued object as an input (such as an adjacency matrix), the method searches over high-dimensional datasets whose observations are split according to a differential condition. Though networks are not directly used, the relational nature of the correlations among high-dimensional datasets implictly reveal network-like structures. The bipartite community detection approach in chapter 5 that correpsonds to the work of Dewaskar et al. (Dewaskar et al., 2020) use a similar logic.

1.2 Contributions to Community Detection

We outline the contributions to the community detetcion and network analysis literature in this section. The two "directions" of community detection algorithms described in this dissertation are:

- *Top-down* approach where memberships and parameter estimates are all already assigned initially and membership probabilities are updated until some global criterion is maximized,

- *Bottom-up* heuristic where test-sets are small at first gradually increase in size until memberships become stable.

Top-down approaches are perhaps more widely used. MOdularity is a pprominent example as the *entire domain* of network combinations are sifted through and then an optimum is chosen. The method of variational inference top estimate stochastic block structures is another example of a top-down approach. We use a form of variational inference known as *hierarchical variational inference* (Ranganath et al., 2016) using hierarchies of memberships in a novel application to networks.

Most existing top-down approaches typically require that all nodes be clustered even if they do not elicit any group structure and necessitate a priori knowledge of the number of groups. The bottom-up approach typically avoids these problems and moreover naturally allows for definition of, and separation between, *signal* and *noise*. Some of these approaches, mentioned in the previouss section, use significance-testing extraction (Wilson et al., 2014; Bodwin et al., 2015; Palowitch et al., 2018). These extraction methods implicitly assume that there is some inherent structure within graphs as dictated by their strengths (sums of weights) and degrees (ie configuration model) but do not assign an explicitly parametric model to these graphs. Furthermore, members not assigned to communities are called *background* nodes (ie. (Palowitch et al., 2018; Wilson et al., 2014)) and are also not statistically modeled.

We note that the iterative testing procedures described by *bottom-up* is not a priori designed to start small and then grow in size (Palowitch et al., 2018). However, in the applications described in this dissertation, the initializing sets are always small or singletons. As such, the analogy in this context is sensible. The testing-based methods proposed in this document are computationally efficient amongst networks with many nodes, edges, and layers, as the mechanism for pattern discovery follow a *bottom-up* heuristic where initializing sets are small and gradually increase in size until memberships become stable (Wilson et al., 2014; Bodwin et al., 2015; Palowitch et al., 2018). . These algorithms contrast with *top-down* approaches which initially posit membership estimates for nodes and iteratively update membership probabilties for each community, or block (Mariadassou et al., 2010; Matias and Miele, 2017). These methods implicitly assume that there is some inherent structure within graphs as dictated by their strengths (sums of weights), degrees, and assortativity structure, but did not assign an explicitly parametric model to these graphs.

Contrasting the algorithmic divide of top-down versus bottom-up, the two overall themes of the *data* catered to the methods proposed by this dissertation are:

1. Detection of communities in multilayer networks that represent different points in time or different modalities

2. Models for irregular networks such as those representing spatial data

The following chapters all reflect one or both of these principles. In Chapters 2, 4 and 5, the methodologies are tailored for networks of many modalities such as multilayer, bipartite, and time series. In Chapters 3- 5, the data used are all from records that represent human mobility or geophysical patterns. Both of these general categories of methods indicate expansions of network-theoretical methods for more modalities of relational data.

1.2.1 Community Detection in Multilayer and Multimodal Networks

While single-graph approaches have been the focus of most work on network theory (Girvan and Newman, 2002; Bender and Canfield, 1978), the literature concerning weighted, multimodal networks is a more recent emerging field of interest (Menichetti et al., 2014; Holme, 2015). The field of *community detection* has, in conjunction, also grown considerably in recent times (Newman, 2018a; Fortunato and Hric, 2016; Handcock et al., 2007; Salter-Townshend and Murphy, 2013). Many techniques have been proposed for *unweighted* (binary) graphs including modularity optimization (Girvan and Newman, 2002; Clauset et al., 2005), stochastic block models (Holland et al., 1983; Nowicki and Snijders, 2001; Peixoto, 2018; Yan et al., 2014), and extraction (Zhao et al., 2012; Lancichinetti et al., 2011). I contribute to the study of multilayer networks primarily through the lens of stochastic blockmodels in this dissertation.

The stochastic block model (SBM) is a foundational theoretical model for random graphs (Karrer and Newman, 2011; Peixoto, 2018; Hoff et al., 2002a; Nowicki and Snijders, 2001) and has also found practical use in community detection (Mariadassou et al., 2010; Newman, 2003). The model lays out a concise formulation for dependency structures within and across communities in networks, but does not typically model *global* characteristics. Though some methods discern but do not statistically model *background* (unclustered) nodes (i.e. Chapters 3-5) (Palowitch et al., 2018;

Wilson et al., 2014; Dewaskar et al., 2020), few existing models explicitly account for *community-wise* noise even though it is useful in many applications. We develop a model for multilayer weighted graphs that explicitly accounts for (1) global noise present between differing communities, and (2) dependency structure across layers within communities. We refer to this model and its associated estimation algorithm as the (multivariate Gaussian) *Stochastic Block (with) Ambient Noise Model* (`SBANM`) for the rest of this manuscript.

Initially proposed to describe binary networks (Hoff et al., 2002a; Nowicki and Snijders, 2001), SBMs have been extended to weighted (Mariadassou et al., 2010)) and multilayer settings (Stanley et al., 2015; Paul and Chen, 2015; Arroyo et al., 2020), and in particular time series (Matias and Miele, 2017) where clusters across all time points have the same inter-block parameters, but varying between-block interactions. These multilayer SBMs typically do not account for correlations between layers. This notion has also only begun to be explored in the context of multilayer SBMs; some recent studies or binary networks have accounted for correlations across layers (Mayya and Reeves, 2019) and noise (Mathews et al., 2019), but typically assume that parameters are already known.

Our main contribution to community detection on multilayer weighted networks is a novel method that jointly finds clusters in a multilayer weighted network and also classifies *what types* of these clusters, namely (local) signal or (global) noise, these are. We propose a (top-down) method that discovers, categorizes, and estimates the associated parameters of these communities. We developed a model as well as its method of inference, which is useful as many existing multilayer SBM analyses assume known parameters (Wang et al., 2019b; Mayya and Reeves, 2019). In the primary case study of Chapter 2 (Chapter 2.5.1), we use `SBANM` to find clusters of diagnostic subgroups of patients judged by similarity measures of their psychopathology symptoms.

In addition, in Chapter 5 I describe joint work with (primarily done by Miheer Dewaskar) to discover clusters in bipartite high-dimensional datasets. We are interested in identifying associations between groups of measured features in two data types in an unsupervised setting that does not make use of auxiliary information about the samples. In particular, the goal is identifying pairs from differing features such that the aggregate (standard Pearson) correlation between features in these features is large.

1.2.2 Demarcating Regions in Spatial Networks

Another contribution of this dissertation is in developing novel community detection methods for networks that represent spatial fields. The method in Chapter 3 is tailored specifically to commuter networks (of human populations) over nodes, which represent spatial points. The highlighted application of the Bimodules Search Procedure (BSP), described in Chapter 5, also uses a method of community detection on bipartite networks to search for clusters that represent spatial locations. In both of these applications, the spatial information is not explicitly accounted for, but the composition of strongly self-looping networks presumes an inherently spatial structure.

Representations of self-loops are inseparable to the notion of distance (spatial or otherwise) in that networks with large self-loops will always be inextricably linked to an implicitly spatial characterization of relational data. Networks with stronger self-loops than cross-edges demonstrate Tobler's "first law of geography" which shows that distal processes that diminish over space. Weighted values associated with such geographical distances are most prominent where there is zero distance (i.e. with itself). In the future directions described in Chapter 6, we outline a method of community detection and analysis of spatial networks that directly takes into account distance information. The distance network may be spatial or social. We also speculate potential dynamics of evolution amongst spatial networks across time, drawing on the basic hypotheses on commuting network formation in Chapter 2.

1.2.3 Finding Clusters in Network Time-Series

Intertemporal community detection similar approach as clustering multilayer networks. Networks in time-series may be seen as a subset of, or a qualitatively different (in this case specifically time-series) multimodal/multilayer networks. Like the `SBANM` method as described in Chapter 2, the intertemporal community detection method finds persistent communities across layers. Persistent communities whose memberships are constant across the entire range of time-points and are perhaps an intrinsic way of describing clusters that span over many time-periods. Such notions intersect heavily with those in developed for multilayer networks in the previous section. In the literature, Liu et al. (Liu et al., 2014) identify and describe the notion of *persistent communities* over multilayer networks, with the goal of distinguishing between steady-state activity and impermanent

behavior. Wilson et al. (Wilson et al., 2016) also proposed a score-based method for multilayer networks with heterogeneous community structure, which captures another class of finding persistent communities across time. On the otehr hand, Stanley et al. and Matias et al. (Stanley et al., 2015; Matias and Miele, 2017), as mentioned in the previous section 1.2.1, use variational inference to find multilayer communities, but do not stipulate the *persistency* requirement i.e. communities may change memberships across time-slices.

The approach described in Chapter 4 is based on directly extracting members from all the nodes in the multilayer network system akin to that the methods of significance testing described in the previous section 1.1.4. The advantages of this method are threefold. Firstly, it is designed for weighted networks of any parametric distribution. Secondly, like the method described in section 1.1.4 as well as in Wilson et al (Wilson et al., 2016), not all nodes are necessarily clustered and some are categorized as background nodes, which capture nodes that are not a part of any community. Thirdly, the method is able to find both "persistent" as well as "heterogeneous" structure. However, the method only applies to networks whose layers represent serial time-slices.

The intertemporal communities discovered from the method proposed in this dissertation serve different descriptive functions for different types of clustering behavior. The intertemporal community detection method described in Chapter 4 looks at weighted networks with registered nodes structured in time series. The communities describe sets of nodes that are connected with each other in varying levels across time. The method is primarily catered towards identifying general directions in the evolution of relationships amongst interconnected nodes. Though the communities are broadly described as increasing, decreasing, or stable in connectivity, neither inter-graph correlation structures, nor specific parameters are not estimated. Moreover, generative forms of such evolutionary structures are only identified in a broad sense but also specified parametrically with means and covariance structures, as well as separation of signal and noise among communities.

The significance-based intertemporal community detection describes and traces general evolutionary patterns in times-series of registered networks without any assumptions in the distributions of weighted edges, but does not make direct inferences about the parameters governing the rates of connectivity, contrasting with `SBANM` as Chapter 2.

1.2.4 Noise in Networks

The notion of *noise* is a central motif in both the major thrusts (*top-down* and *bottom-up*) of community detection methods this dissertation. Noise has always been prevalent in the study of networks but usually relegated to an afterthought partly by design. However, noise as a concept is present and is all of the following chapters and is a major part of Chapter 2. The separation between signal and noise is noted as an advantage in all of the methods preented below. As such, a major contribution of this dissertation is perhaps in the classification and usage of *noise* in network models..

A canonical example of a globally noisy network is the Erdos-Renyi model where every edge is governed by a single probability. The affiliation model is a weighted extension (Allman et al., 2011) used to describe a "noisy homogeneous network"; a single *global* parameter θ_{in} dictates the connectivity between all nodes in *any* community, while another θ_{out} controls the connectivity for all nodes in differing communities. A similar model was posited by Arroyo et al. (Arroyo et al., 2020) where $\theta_{\text{in}} > \theta_{\text{out}}$ as a baseline for network classification. The weighted SBM and the affiliation model are both *mixture models for random graphs* described by (Allman et al., 2011; Ambroise and Matias, 2010). This class of network models accounts for assortativity (the tendency for nodes who connect to each other at similar intensities to cluster) and sparsity (when there are much fewer edges than nodes).

Though there have been many studies on the theoretical properties and empirical usage of SBMs, there have not been much focused on estimating the *noise* inherent within SBMs, much less for multilayer weighted graphs. Extraction-based methods identify background nodes to signify lack of community membership (Palowitch et al., 2018; Wilson et al., 2014), but these methods do not attribute any parametric descriptions to these nodes. Some recent work discuss noise in network models that are oftentimes associated with global (i.e. entire-network) uncertainty that is uniformly added to all edges (Blevins et al., 2021; Newman, 2018b; Mathews et al., 2019; Young et al., 2020). However, few have studied *structural noise* that exists between differing communities or that serves as some notion of a *residual* term (i.e. in regression analysis).

We attempt to address these two gaps in the model proposed in Chapter 2. We simplistically describe the model as follows. In a multilayer graph with Q ground truth communities (indexed

by q), as well as a single block that is considered *noise* (labeled NB for *noise block*), we postulate a model that is *locally unique* with parameter $\boldsymbol{\theta}_q$ for all edges within a block indexed at q. Global noise parameter $\boldsymbol{\theta}_{\text{Noise}}$ describes all interactions between differing blocks as well as NB. This model is written simplistically as follows, but in more detail in Section 2.2:

$$\boldsymbol{\theta}_{ql} = \begin{cases} \boldsymbol{\theta}_q & \text{if } q = l \text{ and } q \text{ is not } NB \\ \boldsymbol{\theta}_{\text{Noise}} & \text{if } q \neq l \text{ or } q \text{ is } NB \end{cases}. \tag{1.1}$$

The model combines qualities of the affiliation model (Allman et al., 2011) with the weighted SBM and extends to multiple layers. Because both the affiliation model and the multilayer SBM are proven to be identifiable by prior work (Allman et al., 2011; Matias and Miele, 2017), we posit that `SBANM` is also identifiable. A brief argument is given in Appendix 2.6.1, but deeper investigation remains as future work. One major advantage of a global noise term is its parsimony compared to SBMs. Existing clustering models on multilayer networks, even when accounting for communities that persist across layers (Liu et al., 2014), still tend toward overparameterization.

A reference or *null* group is often used in scientific and clinical settings, an example being the cerebellum as a reference region-of-interest (ROI) in the analysis of brain networks. The commonality in *out-of-clique* and *baseline* modes of communication in the example in the beginning of Chapter 2 provides an interpretable justification for the empirical realism of this model.

1.3 Outline

We describe the proposed methods in detail starting in Chapter 2, where we introduce a novel class of stochastic blockmodel for multilayer weighted networks that accounts for the presence of a global *ambient* noise governing between-block interactions. We induce a hierarchy of classifications in weighted multilayer networks by assuming that all but one cluster (block) are governed by unique local signals, while a single block behaves identically as interactions across differing blocks (ambient noise). Hierarchical variational inference is employed to jointly detect and typologize blocks as signal or noise. We call this model for multilayer weighted networks the *Stochastic Block (with) Ambient Noise Model* (`SBANM`) and develop an associated community detection algorithm.

Then we apply this method to subjects in the Philadelphia Neurodevelopmental Cohort to discover communities of subjects with similar psychopathological symptoms in relation to psychosis.

In Chapter 3, where we describe a iterative testing method which is an extension of the Continuous Configuration Model Extraction (CCME) method (Palowitch et al., 2018) that finds communities in networks of regional commuting with strongly self looping nodes. Such networks are geographically constrained, but the method does not explicitly make use of spatial features. The method both finds clusters of nodes that are connected in spite of each node posessing predominantly self-looping weights, as well as members (which we call *monads*) that serve as single-node communities.

In Chapter 4, we extend the iterative testing method to temporal networks, and apply it to a variety of human mobility networks that are also implicitly spatially constrained. We introduce new modes of false-discovery rate corrections that account for time-dependent significance testing in networks to support these methods. We identify and interpret the communities that are stably connected across time as well as those that are dynamic and contextualize these with historical trajectories of human mobility flow. Furthermore, we also introduce a method based on anomaly detection that identifies load-imbalanced nodes assuming that they represent carriers of traffic flow (with finite holding capacity).

In Chapter 5, we describe a method for finding bi-clusters amongst two datasets whose clustered members are significantly inter-correlated across the two variable-sets. This method applies to sets of high-dimensional, high-throughput data whose features are correlated across sets. In dealing with modern data, many modes of data analysis, especially exploratory, necessitates the discovery of related feature sets within and across both the datasets. The method is implictly network-based, even if the inputs of the algorithm are not necessarily adjacency matrices but the observations. The method relies on p-values based on permutation distributions that are correspond to sums of squared cross-correlations. Though the method is tailored primarily for genomic datasets, we present an application for identifying zones of related temperature and precipitation in North America.

Finally, in Chapter 6, we describe ongoing and potential directions for future work. This chapter is split into two subsections. The first section of the final chapter details the ongoing and future work branching from `SBANM`. Though not all of the data is yet complete, we describe the integration of imaging, genomic, and cognitive testing data with the survey data as another facet of multilayer

modeling. The methodological goal of this project is to capture all these modes of relational data using multilayer networks and to either (1) classify constellations of symptoms, behaviors, and biomarkers that signify different forms of psychosis and (2) to develop a risk prediction model combining data (whether transformed into networks or not) with the psychosis spectrum status as outcome. In the second part of Chapter 6, we expand on the methodologies of community detection in the networks, as we have previously described. We propose some ideas for theoretical models of the methods described in the following chapters. We also propose methodological extensions for self-looping and temporal networks in previous chapters. Finally, we present ideas for an outline of a null model for spatial networks in the future works section, which is based on the community detection methodology for self-looping networks.

CHAPTER 2

Multivariate Gaussian Blockmodel with Ambient Noise

We describe a model and associated method to find communities with ambient noise in muiltilayer Gaussian weighted networks[1]. We posit an example to motivate the representation of the proposed model to describe patterns in sociality. Suppose there is a social network where nodes represent members and weights represent social interactions. Members naturally interact in cliques where rates of communication are roughly similar (i.e. assortative). Across differing communities, however, rates are assumed to be at a global baseline level. Moreover, interactions among members who are *asocial* and do not belong to any community with a unique signal are similarly modeled as "noise". Who, in a social clique or *scene*, are still friends with each other after 10 years? Alternatively, how might the notion of "friendship" be broken down – in what ways may work relationships (i.e. co-authorships) correlate with social relationships? A schematic figure for this model compared to SBM is presented in Figure 2.1.

The logic of this model is natural for clinical, psychiatric, and experimental settings. Psychiatric illnesses have multiple causes and symptoms. There are no laboratory tests for these conditions. Current diagnostic processes only consider the presence of discrete symptoms and can identify patients who need treatment, but it does not help identify who is at risk for the illness in question. One such illness is schizophrenia, which is a chronic psychotic disorder that affects millions worldwide and imposes a substantial societal burden. Identifying individuals who are at risk for developing this psychotic condition is a clinically significant issue.

In most existing research on networks where nodes represent individuals, edges are known quantities between them. This assumption cannot be applied to psychiatric network models to identify communities of individuals with the same diagnosis as psychiatric disorders manifest with significant heterogeneity. Connections between individuals can be estimated from biological and/or

[1]This chapter is adapted from a manuscript written in 2021 (He et al., 2021) that was joint work with Professors Rose Mary Xavier and Jason Xu

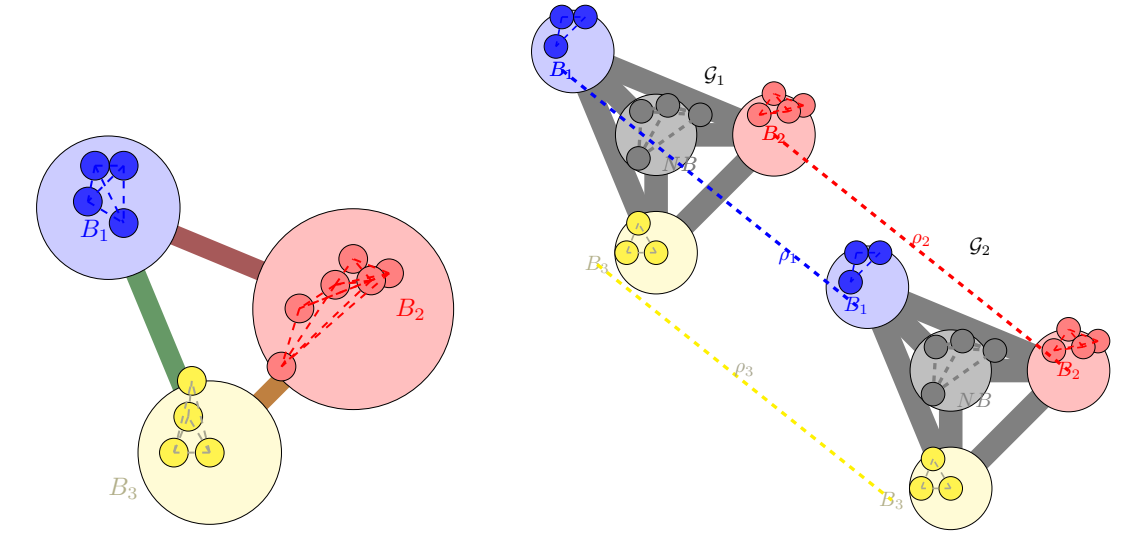


Figure 2.1: Illustrative example of the types of relationships between blocks for the canonical SBM (left) and `SBANM` (right). Dashed lines represent the inter-block connectivity among nodes. Large circles represent distinct communities. Solid thick lines represent the inter-community rates of interaction (transition probabilities if binary). In the canonical case (left), the inter-block transitions are all distinct, as denoted by its colors. For the multilayer `SBANM` case (right), the inter-block parameters are all the same (represented by gray); AN governs the connectivities between blocks and the *intra*-block connectivity within the block NB across two layers $\mathcal{G}_1$ and $\mathcal{G}_2$ with blocks B_1, B_2, B_3 and NB with correlations ρ_1, ρ_2, ρ_3 across $\mathcal{G}_1$ and $\mathcal{G}_2$.

psychosocial data, which can then be used for early identification (Kahn et al., 2015; Clark et al., 1995; Kendell and Jablensky, 2003). With an increase in availability of multimodal data across populations of clinical subjects, multilayer community detection is a natural tool for the classification of psychiatric illnesses with multifaceted characteristics.

While distinguishing psychosis spectrum will be the primary focus of the proposed methodology, it is useful to find latent structure in other types networks. We also demonstrate the method on (1) US congressional voting data and (2) human mobility (bikeshare) data in Sections 2.5.3,2.5.4.We describe the terminology alongside the *Philadephia Neurodevelopmental Cohort* data for the main case study in Section 2.1. We then describe the model and its method of (variational) inference in Section 2.2, and its specific mechanics in Section 2.3. Model performance is assessed and compared with other methods in Section 5. In Section 2.5.1, we demonstrate the focal case study of psychopathology symptom data.

2.1 Data, Notation, and Terminology

For a K-layer **weighted** multigraph with registered n nodes indexed by the set $[n] = \{1, 2, ..., n\}$, let $\mathbf{X}$ represent the collection of multilayer weighted graphs with K layers: $\mathbf{X} = \{\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^K\}$. Similarly, suppose $\mathbf{X}$ contains Q ground truth communities (blocks) indexed by q, but such that a single block is considered *noise* and labeled NB (indexed by q_{NB}). We let $\mathbf{X}_{ij} = (X^1_{ij}, X^2_{ij}, ..., X^K_{ij})$ represent the vector of edge-weights between edges (i, j) across all layers $k = 1, 2, ..., K$. We define a community as $B_q \subset [n]$ to denote the nodes that are contained in a given block indexed by q in $\mathbf{X}$, and we let $\mathbf{X}_q$ represent the set of all edges contained in block q across all K layers:

$$\mathbf{X}_q = \{\mathbf{X}_{ij}\}_{i,j \in B_q}. \tag{2.1}$$

Moreover, we call the set of edges across different blocks q, l (where $q \neq l$) *interstitial noise* (IN), and label it as:

$$\mathbf{X}_{IN} = \{\mathbf{X}_{ij}\}_{i \in B_q, j \in B_l}. \tag{2.2}$$

We fix **one** block indexed as NB as the *noise block* (previously described in Section **??**) where all weights in the block follow a $N_K\big(\boldsymbol{\mu}_{NB}, \boldsymbol{\Sigma}_{NB}\big)$ distribution. This block represents a null region that is devoid of unique signal, but is distributionally governed by the same characteristics as the interstitial relationships between different blocks. We let $\mathbf{X}_{NB}$ represent the set of edges among members in the "noise block": $\mathbf{X}_{NB} = \{\mathbf{X}_{ij}\}_{i,j \in NB}$. In the following subsection we describe the data as introduced in the prior section in the context of the notation. In Section 2.2 we describe the assumption that classifies this notion of noise.

2.1.1 Mapping Notation to Data

Multilayer networks can represent multimodal, longitudinal, or *difference* graphs (Menichetti et al., 2014; Holme, 2015). The data in the *Philadelphia Neurodevelopmental Cohort* (PNC) (described below) is constructed as a multimodal network, while the applications outlined in Appendices 2.5.3 and 2.5.4 are examples of longitudinal graphs. In each application, we write the weighted graph-system $\mathbf{X}$ with K layers and define the index set $[n]$ as the set (of cardinality n) of all nodes. Every layer has n nodes and each weight $\mathbf{X}_{ij}$ between nodes i, j is written as a $K-$ dimensional vector, and each layer-specific (at k) weight is written as X_{ij}^k.

With respect to the PNC data, $\mathbf{X}$ represents the whole set of anxiety, behavior, and mood psychopathology symptom networks across a given set of subjects. There are three layers $\mathbf{X}^x, \mathbf{X}^y, \mathbf{X}^z$ indexed by $k = \{x, y, z\}$; each represents one of the psychometric evaluation networks representing each disorder. The sample size n in this context represents the 5136 subjects between the ages of 11 to 17 (*youth*) and 1863 between the ages of 18 to 21 (*early adult*). Each node represents a subject, and each weighted edge the transformed similarity ratio between two subjects for anxiety, behavior, and mood symptoms.

A community sample was obtained from the PNC study from the greater Philadelphia area. Subjects aged 8-21 years were subject to a detailed neuropsychiatric evaluation (Calkins et al., 2014, 2015). This sample is used as the primary case study. X_{ij}^k is assumed to be generated from clusters of nodes whose (Fisher) transformed edges follow blockwise multivariate normal distributions. We use three general categories of disorders to represent each layer:

1. Anxiety ($\mathbf{X}^x$): 44 questions (generalized, social, separation anxiety, etc.)

2. Behavior ($\mathbf{X}^y$): 22 questions (ADHD, OCD, CDD)

3. Mood ($\mathbf{X}^z$): 10 questions (depression and mania),

then Fisher-transformed to produce the weighted edge in graph $\mathbf{X}^k$ in layer k. In these following sections these categories will simply be referred to as "anxiety","behavior", and "mood". More details on pre-processing can be found in Section 2.1.2.

2.1.2 PNC Preprocessing and Network Construction

The PNC has a well-represented sample with mostly European American ancestry but a substantial portion of African Americans. Roughly 21% met psychosis spectrum criteria, 4% reported threshold psychosis symptoms, 12% reported subthreshold positive symptoms, 2% exhibited subthreshold negative symptoms ((Calkins et al., 2017)). We separately analyze the two age cohorts *youth* (with sample size 5136) and *early adult* (sample size 1863).

Response networks are constructed using a function that gauges similarity as well as positivity or negativity of responses. This distance function is similar to Hamming distance, but takes into account the direction of *positive* or *negative* agreement and is between -1 and 1 . In a single graph-layer $\mathbf{X}^k$, a weight X^k_{ij} between two nodes is derived from indicators $h^k_{ij,u}$ across U questions (indexed by u) pertaining to a given set of conditions.

$$h^k_{ij,u} = \begin{cases} 1 & \text{if } i,j \text{ both answer "yes"} \\ -1 & \text{if } i,j \text{ both answer "no"} \\ 0 & \text{otherwise} \end{cases}$$

Each $h^k_{ij,u}$ between two subjects u, v is -1 if both answer no, 1 if both yes, otherwise 0. These values are then summed and divided by the total number of questions U:

$$r^k_{ij} = \frac{\sum_{u=1,..,U} h^k_{ij,u}}{U}.$$

The weight r^k_{ij} is 1 if two subjects both answer yes to everything and -1 if they answer no to everything. The weight r^k_{ij} is then transformed using a Fisher transformation to produce a value that approximates an observation in a normal distribution, in layer k: $X^k_{ij} = \text{Fisher}(r^k_{ij})$.

2.2 Model and Inference

`SBANM` supposes that networks across K layers have the same block structure, while transition parameters between blocks are fixed at the same global, *ambient*, level. This model allows detection of common latent characteristics across layers, as well as differential sub-characteristics within blocks (represented by multivariate normal distributions). This model also presumes block structures whose edges are correlated across layers.

Definition 2.1. (Correlated Signal Blocks) For a $K-$layer (Gaussian) weighted multigraph $\mathbf{X} = \{\mathbf{X}^1, ..., \mathbf{X}^K\}$ where each layer k represents a graph with n registered nodes , let $B_q \subset [n]$ represent a community housing a partition of nodes $\{i\}_{i \in B_q}$, then each weighted edge between any node in block B_q form a multivariate normal distribution with mean K-dimensional vector $\boldsymbol{\mu}_q$ and $K \times K$-dimensional covariance matrix $\boldsymbol{\Sigma}_q$:

$$\boldsymbol{\Sigma}_q = \begin{pmatrix} \sigma^2_{q,1} & \rho_q \sigma_{q,1}\sigma_{q,2} \cdots & \rho_q \sigma_{q,1}\sigma_{q,K} \\ \rho_q \sigma_{q,2}\sigma_{q,1} & \sigma^2_{q,2} \cdots & \rho_q \sigma_{q,2}\sigma_{q,K} \\ \cdots & \cdots & \cdots \\ \rho_q \sigma_{q,K}\sigma_{q,1} & \cdots & \sigma^2_{q,K} \end{pmatrix}.$$

If nodes i, j are in the same block, the distribution of their edges follow a multivariate normal distribution

$$\mathbf{X}_{ij} | \{i \in B_q, j \in B_q\} \sim N_K(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q).$$

Note that there is a single correlation parameter ρ_q across all layers for a given block B_q, implying that $\boldsymbol{\Sigma}_q$ is a rank 1 plus rank 2 matrix. This is a deliberate choice to induce parsimony and interpretability among block relationships across all layers. We assume that the *noise block* as has the same characteristics as the *interstitial noise*; both are drawn from the same distribution AN (*ambient noise*). AN is a global noise distribution that governs both IN and NB:

$$\mathbf{X}_{IN} \overset{d}{=} \mathbf{X}_{NB} \sim N_K(\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}).$$

Because NB and IN both represent "baseline" levels of connectivity for the network, we assume that they both have equivalent characteristics as AN. Members of each block B_q interact with other members in the same block at rates that follow multivariate $\boldsymbol{\mu}_q$ with variance $\boldsymbol{\Sigma}q$, but interact with members in differing groups $l; l \neq q$ at baseline rates $\boldsymbol{\mu}_{IN}$ with variance $\boldsymbol{\Sigma}_{IN}$, i.e. background interactions.

Definition 2.2. (Ambient Noise) Edges in IN between differing blocks and in NB, are characterized by $(\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})$: $\boldsymbol{\Sigma}_{AN}$ is a $K \times K$ diagonal matrix with diagonal $(\sigma^2_{AN,1}, ..., \sigma^2_{AN,K})$ and off-diagonal entries of 0:

$$\mathbf{X}_{ij} | \{i \in B_q, j \in B_l\} \sim N_K(\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}).$$

For a community $B_q \subset [n]$ representing the nodes that are contained in block q in a weighted multilayer network $\mathbf{X}$,we let $\mathbf{X}_q$ represent the set of all edges contained in block B_q across all K layers as defined in Equation (2.1). Conversely, the set of edges across differing B_q, B_l (i.e. interstitial noise), are defined as in Equation (2.2).

Definition 2.3. (Stochastic Block (with) Ambient Noise Model (`SBANM`)) A $K-$layer (Gaussian) weighted multigraph $\mathbf{X} = \{\mathbf{X}^1, ..., \mathbf{X}^K\}$ with n nodes and Q communities (blocks) indexed by q with a single block that is considered *noise* labeled NB (indexed by q_{NB}) with disjoint blocks $\{B_1, B_2, ..., NB, ..., B_Q\}_{q:q \leq Q}$ such that $\bigcup_{q \leq Q} B_q \bigcup NB = [n]$ is a `SBANM` if the following conditions are satisfied.

1. Edges in the same block B_q adhere to (Correlated Signal Blocks) where each edge $\mathbf{X}_{ij}$ follows conditional distribution $N_K(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ given block memberships,

2. Ambient noise AN with $N_K(\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})$ governs both IN and NB:

   (a) Edges $i \in B_q$ and $j \in B_l$ $(l \neq q)$ follow a $N_K(\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})$ distribution.

   (b) **One** block NB contains members whose edges are generated from a $K-$ dimensional multivariate normal distribution $N_K(\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})$.

2.2.1 Connection to Existing Models

The weighted SBM and the affiliation model are both cases of the *mixture models for random graphs* described by Allman et al. (Allman et al., 2011; Ambroise and Matias, 2010). This general class of network models accounts for assortativity (the tendency for nodes who connect to each other at similar intensities to cluster together) and sparsity (when there are much fewer edges than nodes). In addition to the class of VEM-based inference methods (Mariadassou et al., 2010; Matias and Miele, 2017; Paul and Chen, 2018) that are extensively referenced in Section 1.2.4, we also note multigraph SBM inference methods based on spectral decomposition (Wang et al., 2019b; Arroyo et al., 2020; Mayya and Reeves, 2019). These methods are typically applied to binary networks and use different sets of methodology or assumptions such as known parameters ((Mayya and Reeves, 2019)), but are still similar in motivation as to warrant comparison. Some of these existing methods model edge connectivity of a (potentially multilayer) network as a function of membership vectors $\mathbf{Z}_i$ (for node i), connectivity matrix $\mathcal{R}_k$ at layer k, and the graph Laplacian (Mayya and Reeves, 2019; Mathews et al., 2019; Reeves et al., 2019; Arroyo et al., 2020; Wang et al., 2019b). Typically, the connectivity rate corresponds to Bernoulli probabilities (for binary networks), but some of these approaches allow for (or posit for future work) extensions to the weighted cases (Wang et al., 2019b; Mercado et al., 2019; Arroyo et al., 2020). Some work has focused on studying the correlations or linear combinations of the eigenvectors of $\mathcal{R}_k$, but in most of these cases *conditional independence given labels* between layers is assumed for correlated networks (Mayya and Reeves, 2019; Arroyo et al., 2020). A recent trend in these multiplex methods has focused on devising an optimal aggregation scheme to combine multiple layers and then to use single-graph methods on the resultant static network (Levin et al., 2019). We consider several special cases for `SBANM` where it reduces to existing models.

1. If all ρ_q were zero (ie. diagonal $\mathbf{\Sigma}_q$; no correlations amongst communities) and all the within-community signals were the same, then `SBANM` is a multivariate extension of the models posited by Allman et al. (Allman et al., 2011) or Arroyo et al. (Arroyo Relión et al., 2019).

2. If $K = 1$, `SBANM` is a special case of the weighted Gaussian SBM as proposed by Mariadassou et al. where all inter-block connectivities are fixed at a single rate (Mariadassou et al., 2010).

3. Wang et al. ((Wang et al., 2019b)) constrain the connectivity matrix to a diagonal, which would be analogous to `SBANM` if ambient noise parameter is fixed at zero: $\boldsymbol{\theta}_{AN} := 0$.

4. Arroyo et al. (Arroyo et al., 2020) describe the multilayer SBM (Holland et al., 1983) for binary graphs which "could be easily extended to the weighted cases". The model assumes *independent* block parameters $\mathcal{R}_k$ across every layer. If there were parameters $\boldsymbol{\theta}_{AN}$ such that $\mathcal{R}_{ql,k} := \boldsymbol{\theta}_{AN}$ (for every $q \neq l$), then a special case of `SBANM` (where each $\rho_q := 0$) would be recovered.

`SBANM` finds a loose connection to mixed-membership blockmodels (MMBM) in that both models attribute uncertainty to membership designations (Airoldi et al., 2007). However, MMBM doubly complicates the model parameter landscape with overlapping block combinations, while `SBANM` more parsimoniously addresses ambiguous memberships by subsuming their characteristics into an umbrella ambient noise term that describes the ambiguities in block memberships in the interstitial noise term IN.

2.2.2 Variational Inference

The proposed model is estimated by variational inference (VI) , which has historically been used for estimating SBM memberships as well as their parameters (Mariadassou et al., 2010; Paul and Chen, 2015). VI is an approach to approximate a conditional density of latent variables using observed information by solving this problem with optimization (Blei et al., 2017; Jaakkola, 2000). When optimizing the full likelihood is intractable, simpler surrogates of complicated variables are chosen as to create a simpler objective function. The Kullback-Liebler (KL) Divergence between this simpler function and the full likelihood are then minimized. For community detection problems, mean-field (MF) approximations of membership allocations often serve as simpler surrogates of latent approximands to simplify the likelihood function into a lower bound (typically known as *evidence lower bound*: ELBO) (Mariadassou et al., 2010; Salter-Townshend and Murphy, 2013; Airoldi et al., 2007).

Variational EM (VEM) is the state-of-the-art for SBM estimation and demonstrably more efficient than other approaches (such as MCMC) (Mariadassou et al., 2010; Nowicki and Snijders, 2001). Daudin et al. introduced using VEM for binary-graph SBMs ((Daudin et al., 2008). Mari-

adassou et al. used a similar method for detecting communities in a single weighted graph (Mariadassou et al., 2010), while Matias et al. also did so for multilayer networks (Matias and Miele, 2017). The estimation algorithm for the proposed model is also rooted in VEM, but we augment the procedure with *signal* and *noise* typologizing different blocks.

Though it enables efficient inference, "typical" MF VI is limited by its assumption of strong factorization and does not capture posterior dependencies between latent variables arising amongst multilayered networks. Hierarchical Variational Inference (HVI) provides a natural framework for the two-layered latent structure for multilayer networks. A natural hierarchy is induced in `SBANM` by the assumption that all but one block are under the umbrella of *signal*, while a single block is classified as noise. HVI augments variational approximations with priors on its parameters: this assumption allows joint clustering of blocks and their signal-noise differentiation as the *superstructure*.

We use a similar approach to that originally used in Daudin et al. (Daudin et al., 2008). The latent variable of interest is the membership allocation matrix $\mathbf{Z}$, which is a $n \times Q$ matrix where each row $\{\mathbf{Z}_i\}_{i \leq n}$ contains $Q - 1$ zeros and a single 1 that represents membership at that given entry. We introduce indicator $\mathbf{C}$ to determine if a block q is signal or noise NB. $\mathbf{C}$ is a vector of length Q whose values C_q are 0 or 1. The main difference between our's and previous approaches is that joint approximate conditional distributions of $\mathbf{Z}$ and $\mathbf{C}$ are modeled instead of just $\mathbf{Z}$:

$$R_{\mathbf{X}}(\mathbf{Z}, \mathbf{C}) \approx \prod_{i,q} \Big(m(\mathbf{Z}_i, \boldsymbol{\tau}_i) \times \text{Bern}(C_q, P_q) \Big). \tag{2.3}$$

In Eq. (2.3) $R_{\mathbf{X}}(\mathbf{Z}, \mathbf{C})$ represents the joint variational distribution of the memberships $\mathbf{Z}, \mathbf{C}$. The exact joint distribution is unknown, but the hierarchical mean field (MF) approximation $R(\mathbf{Z}, \mathbf{C})$ can be used to obtain a factorized estimate for its marginals (Ranganath et al., 2016). We write the approximate composition of marginals using "$\times$"; $m(\cdot)$ represents the multinomial distribution. The variational approximations of membership matrix $\mathbf{Z}$ is a $n \times Q$-dimensional matrix $\boldsymbol{\tau}$, each row represents the vector of probabilities that approximates $\mathbf{Z}_i$ (Mariadassou et al., 2010).

The variational approximation of the indicator C_q at block q is the probability P_q, which typologizes (and "sits at a higher tier" than) $\boldsymbol{\tau}$. Under variational distribution R, each member i of a block B_q adheres to multinomial distribution with parameter $\tau_{iq} = \mathbb{E}[\mathbf{Z}_{iq}]$. P_q is the probability

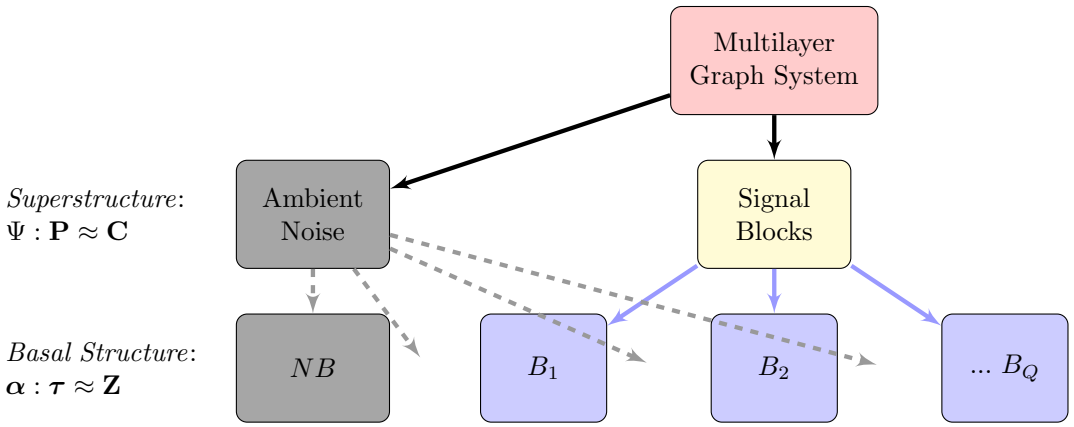


Figure 2.2: Schematic diagram for the hierarchy of organization for blockstructures with signal/noise differentiation for blocks as the top layer and the actual blocks as the bottom layer.

of C_q akin to τ_{iq}. For each q, P_q is ambient noise with prior probability Ψ. A derivation for Ψ is given in Section 2.3.5

Definition 2.4. Ψ is the probability of block $\{B_q\}_{q:q\leq Q}$ to be noise block NB:

$$\Psi := (Q-1)/Q \tag{2.4}$$

The hierarchical MF distribution $R_{\text{hv}}(\mathbf{Z})$ as introduced by Ranganathan et al. (Ranganath et al., 2016) "marginalizes out" the MF parameters in $R_{\mathbf{X}}(\mathbf{Z}, \mathbf{C})$ and is written as

$$R_{\text{hv}}(\mathbf{Z}) = \int R_{\mathbf{X}}(\mathbf{Z}, \mathbf{C}) \mathrm{d}\mathbf{C}.$$

Following the methods of estimation proposed in prior work on SBM estimation (Daudin et al., 2008; Mariadassou et al., 2010; Paul and Chen, 2018), $R_{\mathbf{X}}(\mathbf{Z}, \boldsymbol{\tau})$ represents the multinomial variational distribution wherein each τ_{iq} approximates the membership allocations. The integrated $R_{\text{hv}}(\mathbf{Z})$ represents the distribution in prior work that is not subject to the signal or noise categorizations.

2.2.3 ELBO and Hierarchical ELBO

This section describes the hierarchical ELBO as well as the derivations for these expressions. Prior VEM-based estimation methods focus on optimizing the Evidence Lower Bound (ELBO) (Paul and Chen, 2015, 2018; Mariadassou et al., 2010; Daudin et al., 2008). $\mathcal{L}$ is the approximately

optimal likelihood that minimizes the KL Divergence between $R(\mathbf{Z}, \mathbf{C})$ and the posterior frequency $f(\mathbf{Z}, \mathbf{C}|\mathbf{X})$. It is the sum of the expected frequency and the entropy $\mathcal{H}$ of variational variable $\mathbf{Z}$:

$$\mathcal{L} = \mathbb{E}_{R_{\text{hv}}(\mathbf{Z})}[\log f(\mathbf{Z}, \mathbf{X})] + \mathcal{H}_{\text{hv}}(R(\mathbf{Z})).$$

A better bound than the ELBO is derived by introducing the marginal recursive variational approximation $S(\mathbf{C}|\mathbf{Z})$, and then exploiting the following inequality with joint MF distribution $R(\mathbf{Z}, \mathbf{C})$ and the (hierarchical) entropy $\mathcal{H}(\mathbf{Z})$:

$$\mathcal{H}_{\text{hv}}(R(\mathbf{Z})) \geq -\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\left[\log R(\mathbf{Z}, \mathbf{C})\right] + \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\left[\log S(\mathbf{C}|\mathbf{Z})\right]. \tag{2.5}$$

A proof of inequality (2.5) is given as follows (Ranganath et al., 2016) : An inequality can be drawn between the "ordinary" ELBO $\mathcal{L}$ without any hierarchical information and the Hierarchical ELBO

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{R_{\text{hv}}(\mathbf{Z})}[\log f(\mathbf{Z}, \mathbf{X})] + \mathcal{H}_{\text{hv}}(R(\mathbf{Z})) \\
&\geq \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{Z}, \mathbf{X})] - \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\left[\log R(\mathbf{Z}, \mathbf{C})\right] + \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\left[\log S(\mathbf{C}|\mathbf{Z})\right] \\
&:= \mathcal{L}'(\text{Hierarchical ELBO }).
\end{aligned}$$

The inequality in the above relationship arises from the decomposition of the entropy $\mathcal{H}_{\text{hv}}$ of the hierarchical distribution. The proof of the inequality is based on the proof from Ranganath et al. (Ranganath et al., 2016) :

Proposition 1.

$$\mathcal{H}_{hv}(R(\mathbf{Z})) \geq -\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\left[\log R(\mathbf{Z}, \mathbf{C})\right] + \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\left[\log S(\mathbf{C}|\mathbf{Z})\right].$$

Proof.

$$
\begin{aligned}
\mathcal{H}_{\text{hv}}(R(\mathbf{Z})) &= -\mathbb{E}_{R_{\text{hv}}(\mathbf{Z})}[\log R_{\text{hv}}(\mathbf{Z})] \\
&= -\mathbb{E}_{R_{\text{hv}}(\mathbf{Z})}\left[\log R_{\text{hv}}(\mathbf{Z}) - \mathbf{KL}\left(R_{\mathbf{C}|\mathbf{Z}}(\mathbf{C}|\mathbf{Z}); R_{\mathbf{C}|\mathbf{Z}}(\mathbf{C}|\mathbf{Z})\right)\right] \\
&\geq -\mathbb{E}_{R_{\text{hv}}(\mathbf{Z})}\left[\log R_{\text{hv}}(\mathbf{Z}) + \mathbf{KL}\left(R_{\mathbf{C}|\mathbf{Z}}(\mathbf{C}|\mathbf{Z}); S(\mathbf{C}|\mathbf{Z})\right)\right] \\
&= -\mathbb{E}_{R_{\text{hv}}}\left[\mathbb{E}_{R(\mathbf{Z})}[\log R_{\text{hv}}(\mathbf{Z})] + \log R_{\mathbf{C}|\mathbf{Z}}(\mathbf{C}|\mathbf{Z}) - \log S(\mathbf{C}|\mathbf{Z})\right] \\
&= -\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\left[\log R_{\text{hv}}(\mathbf{Z}) + \log R_{\mathbf{C}|\mathbf{Z}}(\mathbf{C}|\mathbf{Z}) - \log S(\mathbf{C}|\mathbf{Z})\right] \\
&= -\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\left[\log R_{\mathbf{Z},\mathbf{C}}(\mathbf{Z},\mathbf{C}) - \log S(\mathbf{C}|\mathbf{Z})\right]
\end{aligned}
$$

□

The jointly factorized MF components of $R(\mathbf{Z},\mathbf{C}) = R(\mathbf{C})R(\mathbf{Z}|\mathbf{C})$ are written as follows: $R(\mathbf{C}) = \prod_q P_q^{C_q}(1-P_q)^{1-C_q}$ as each C_q is Bernoulli distributed, and $R(\mathbf{Z}|\mathbf{C})$ is written similarly to prior variational membership variables (Daudin et al., 2008; Mariadassou et al., 2010), exponentiated by C_q:

$$
R(\mathbf{Z}|\mathbf{C}) = \prod_q \prod_i \left(\tau_{iq}^{Z_{iq}}\right)^{C_q} \left(\prod_i \tau_{iq}^{Z_{iq}}\right)^{1-C_q},
$$

combining to form $R(\mathbf{Z},\mathbf{C}) = R(\mathbf{Z}|\mathbf{C})R(\mathbf{C})$. Moreover, the recursive variational approximation $S(\mathbf{C}|\mathbf{Z})$, as introduced by Ranganath et al. (Ranganath et al., 2016) estimates the higher-order memberships $\mathbf{C}$ using the basal memberships $\mathbf{Z}$:

$$
S(\mathbf{C}|\mathbf{Z}) = \prod_q \prod_i \left(\Psi^{C_q}(1-\Psi)^{1-C_q}\right)^{Z_{iq}}.
$$

The global signal rate $\Psi := \mathbb{P}(NB)$ (Definition 2.4) represent the *prior* probabilties of each group membership C_q, or the parameters of the *initial stationary distribution* of P_q (Matias and Miele, 2017).

Definition 2.5. (Evidence Lower Bound (ELBO)) Given observed data $\mathbf{X}$ with unknown latent membership variables $\mathbf{Z}$, the evidence lower bound (ELBO) $\mathcal{L}$ is the approximately optimal likelihood that minimizes the KL Divergence between the approximate distribution $R(\mathbf{Z},\mathbf{C})$ and the

posterior frequency $f(\mathbf{Z}, \mathbf{C}|\mathbf{X})$. It is expressed as follows:

$$\mathcal{L} = \mathbb{E}_{R_{\text{hv}}(\mathbf{Z})} \left[\log f(\mathbf{Z}, \mathbf{X}) - \log R_{\text{hv}}(\mathbf{Z})\right]$$

Alternatively, the ELBO can be rewritten as the sum of the expected frequency and the entropy $\mathcal{H}$ of variational variable $\mathbf{Z}$:

$$\mathcal{L} = \mathbb{E}_{R_{\text{hv}}(\mathbf{Z})}[\log f(\mathbf{Z}, \mathbf{X})] + \mathcal{H}_{\text{hv}}(R(\mathbf{Z})).$$

We write $\mathcal{L}'$ here as follows:

$$\mathcal{L}' = \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{Z}, \mathbf{X})] - \mathbb{E}_{R(\mathbf{Z},\mathbf{C})} \left[\log R(\mathbf{Z}, \mathbf{C})\right] + \mathbb{E}_{R(\mathbf{Z},\mathbf{C})} \left[\log S(\mathbf{C}|\mathbf{Z})\right]$$

The log likelihood portion of the hierarchical ELBO is written as :

$$\begin{aligned}
\mathbb{E}_{R_{\boldsymbol{X}}}[\log(f(\mathbf{X}|\mathbf{Z}))] &= \sum_q P_q \sum_i \sum_j \ \tau_{iq}\tau_{jq}\left(\frac{1}{2}(\mathbf{X}_{ij} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q^{-1}(\mathbf{X}_{ij} - \boldsymbol{\mu}_q) - (2\pi)^{K/2}(\log|\boldsymbol{\Sigma}_q|)^{1/2}\right) \\
&+ \sum_q (1 - P_q) \sum_i \sum_j \tau_{iq}\tau_{jq}\left(\frac{1}{2}(\mathbf{X}_{ij} - \boldsymbol{\mu}_{AN})^T \boldsymbol{\Sigma}_{AN}^{-1}(\mathbf{X}_{ij} - \boldsymbol{\mu}_{AN}) - (2\pi)^{K/2}(\log|\boldsymbol{\Sigma}_{AN}|^{1/2}\right) \\
&+ \sum_q \sum_{l:l\neq q} \sum_i \sum_j \ \tau_{iq}\tau_{jl}\left(\frac{1}{2}(\mathbf{X}_{ij} - \boldsymbol{\mu}_{AN})^T \boldsymbol{\Sigma}_{AN}^{-1}(\mathbf{X}_{ij} - \boldsymbol{\mu}_{AN}) - (2\pi)^{K/2}(\log|\boldsymbol{\Sigma}_{AN}|^{1/2}\right).
\end{aligned}$$

The full form of the hierarchical ELBO is the log likelihood part plus the membership probabilities, entropy, and their hierarchical counterparts is shown as follows:

$$\begin{aligned}
\mathcal{L}' = \mathbb{E}_{R_{\boldsymbol{X}}}[\log(f(\mathbf{X}|\mathbf{Z}))] &+ \sum_{i,q} \tau_{iq} \log \alpha_q - \sum_q \sum_i \tau_{iq} \log \tau_{iq} - \\
&\sum_q \Big(P_q \log P_q + (1 - P_q)\log(1 - P_q)\Big) + \sum_i \sum_q \Big(P_q \log \Psi + (1 - P_q)\log(1 - \Psi)\Big)\tau_{iq}
\end{aligned}$$

This is the full expression for the hierarchical ELBO as described in Section 2.2.5.

2.2.4 Parameter Estimation

In this section we describe the estimation of parameters. First, to ease notation, we introduce some more terms

$$f(X_{ij}^k, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) = \frac{1}{2}(X_{ij}^k - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q^{-1}(X_{ij}^k - \boldsymbol{\mu}_q) - (2\pi)^{K/2}(\log|\boldsymbol{\Sigma}_q|)^{1/2} \tag{2.6}$$

$$f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}) = \frac{1}{2}(X_{ij}^k - \boldsymbol{\mu}_{AN})^T \boldsymbol{\Sigma}_{AN}^{-1}(X_{ij}^k - \boldsymbol{\mu}_{AN}) - (2\pi)^{K/2}(\log|\boldsymbol{\Sigma}_{AN}|)^{1/2}. \tag{2.7}$$

Equation (2.6) denotes the density for edges in a signal block $\left(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right)$ at layer k; equation (2.6) denotes density for edges with noise $\left(\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}\right)$. In a graph $\mathbf{X}$ with K graph-layers $\{\mathbf{X}^1, ..., \mathbf{X}^K\}$, each edge between nodes i, j of each layer k has conditional density

$$\begin{aligned}\log f(\mathbf{X}|\mathbf{Z}) = &\sum_{q:B_q \neq NB; q \leq Q} \sum_{k \leq K} \sum_{i,j \leq n} \tau_{iq}\tau_{jq} f(X_{ij}^k, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \\ &+\mathbf{1}(B_q = NB) \sum_{i,j \leq n} \tau_{iq}\tau_{jl} f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}) + \sum_{q,l \leq Q: q \neq l} \sum_{i,j \leq n} \tau_{iq}\tau_{jl} f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}).\end{aligned} \tag{2.8}$$

The log likelihood portion of the ELBO, $\log(f(\mathbf{X}|\mathbf{Z}))$, written above in Equation (2.8) is comprised of three parts: unique signals for every q (top), the noise block NB (bottom left), and the interstitial noise IN (bottom right). AN is the global *ambient noise* whose parameters govern the *interstitial noise* as well as *noise block* as in Definition 2.2. The probability of block B_q "being signal" is demarcated by P_q. Given variational variables $\boldsymbol{\tau}, \mathbf{P}$, the expected likelihood is

$$\begin{aligned}\mathbb{E}_{R_{\boldsymbol{X}}}[\log(f(\mathbf{X}|\mathbf{Z}))] = &\sum_{q:q \leq Q} \mathbb{P}(B_q \neq NB)\tau_{iq}\tau_{jl} f(X_{ij}^k, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \\ &+\mathbb{P}(B_q = NB) \sum_{i,j:i \neq j} \tau_{iq}\tau_{jl} f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}) + \sum_{q,l \leq Q: q \neq l} \sum_{i,j \leq n} \tau_{iq}\tau_{jl} f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}).\end{aligned}$$

The $\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{Z})]$ term restores to the same form as earlier work on SBMs (Mariadassou et al., 2010; Daudin et al., 2008):

$$\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{Z})] = \sum_{i,q} \tau_{iq} \log \alpha_q, \tag{2.9}$$

where as in prior work (Daudin et al., 2008; Mariadassou et al., 2010), the variables α_q represent the membership probabilities of Z_{iq} and sum to 1:

$$\alpha_q = \mathbb{P}(i \in B_q) = \mathbb{P}(Z_{iq} = 1). \tag{2.10}$$

Here we show that the term for $\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{Z})]$ as written in Eq. (2.9) is the same as in prior studies such as Daudin et al. (Daudin et al., 2008)

Proposition 2.

$$\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{Z})] = \sum_{i,q} \tau_{iq} \log \alpha_q$$

Proof.

$$\begin{aligned}
\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{Z})] &= \sum_i \sum_q \Big(P_q \tau_{iq} \log \alpha_q + (1 - P_q) \tau_{iq} \log \alpha_q \Big) \\
&= \sum_q (P_q + (1 - P_q)) \Big(\sum_i \tau_{iq} \log \alpha_q \Big) \\
&= \sum_{i,q} \tau_{iq} \log \alpha_q.
\end{aligned}$$

□

Note for the rest of the manuscript we use to $\sum_{i,q}(\cdot)$ to signifiy the double summation across all $i \leq n$ and $q \leq Q$. Since the expected log frequency of the membership vectors $\mathbf{Z}$ reduces to that in canonical SBMs. The joint density is written as:

$$\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{X}, \mathbf{Z})] = \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{X}|\mathbf{Z})] + \sum_{i,q} \tau_{iq} \log \alpha_q. \tag{2.11}$$

Model parameters can be partitioned into Θ_{Signal} and Θ_{Noise} in addition to global parameters $\boldsymbol{\alpha}, \Psi$. We write the entire set of model parameters as

$$\Theta = \{\boldsymbol{\alpha}, \Psi, \Theta_{\text{Noise}}, \Theta_{\text{Signal}}\}. \tag{2.12}$$

$\Theta_{\text{Signal}} = \{\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\}_{q:1 \leq q \leq Q; B_q \neq NB}$ represents the model parameters that are unique to each block B_q (not including NB), and also there is one fixed label q_{NB} that indexes the noise block NB.

$\Theta_{\text{Noise}} = \{\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}\}$ represents the noise parameters that govern both interstitial noise IN and noise block NB. For NB, each correlation between layers is set at zero.

2.2.5 Decomposition of the Hierarchical ELBO

The estimation procedure optimizes the hierarchical ELBO. The hierarchical ELBO $\mathcal{L}'$ (details in Appendix 2.2.3) can be decomposed as

$$\mathcal{L}' = \mathbb{E}_{R(\mathbf{Z},\mathbf{C})} \log f(\mathbf{X}, \mathbf{Z}) + \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log R(\mathbf{C}, \mathbf{Z})] + \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log S(\mathbf{C}|\mathbf{Z})]. \tag{2.13}$$

The first term $\mathbb{E}_{R(\mathbf{Z},\mathbf{C})} \log f(\mathbf{X}, \mathbf{Z})$ which represents the observed joint densities of $\mathbf{X}$ and $\mathbf{Z}$ is written in Eq. (2.11). $\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log R(\mathbf{C}, \mathbf{Z})]$ represents the joint distribution of the two-tiered variational variables and is written as:

$$\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log R(\mathbf{Z}, \mathbf{C})] = \sum_{i,q} \tau_{iq} \log \tau_{iq} + \sum_{q} \Big(P_q \log P_q + (1 - P_q) \log(1 - P_q) \Big).$$

The third term $\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log S(\mathbf{C}|\mathbf{Z})]$ described by Ranganath et al. as the 'recursive variational approximation' (Ranganath et al., 2016) for $R(\cdot)$, is written as

$$\mathbb{E}_{R(\mathbf{Z},\mathbf{C})} \log S(\mathbf{C}|\mathbf{Z})) = \sum_{i,q} \Big(P_q \log \Psi + (1 - P_q) \log(1 - \Psi) \Big) \tau_{iq}.$$

Combining these elements together, the hierarchical ELBO can be rewritten as:

$$\begin{aligned}\mathcal{L}' =& \mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{X}|\mathbf{Z})] + \sum_{i,q} \Big(\tau_{iq} \log \alpha_q + \tau_{iq} \log \tau_{iq} + \Big(P_q \log \Psi + (1 - P_q) \log(1 - \Psi) \Big) \tau_{iq} \Big) \\ &+ \sum_{q} \Big(P_q \log P_q + (1 - P_q) \log(1 - P_q) \Big).\end{aligned}$$

The hierarchical ELBO written in full can be found in Section 2.2.3. Derivations for all of these terms can be found in the following sections.

2.3 Estimation Algorithm

We summarize the targets of inference here to set up the language for the rest of the section. We distinguish *variational* and *model parameters*: variational parameters $\boldsymbol{\tau}_q$ and P_q (for $q : q \leq Q$) approximate the membership allocations, while model parameters describe the parametric qualities of the blocks. Within the set of model parameters, we further distinguish *local* and *global* parameters. *Local* parameters are $\boldsymbol{\Sigma}_q$, and membership probabilties α_q for each q. *Global* parameters are $\Psi, \Theta_{\text{Noise}}$. We use VEM to estimate variational parameters in the E-step and model parameters in the M-step, alternating these steps until the differences in $\boldsymbol{\tau}$ become miniscule. Operationally, the E-step and M-step are implemented in a an alternating fashion until the membership variables $\boldsymbol{\tau}$ reach some criterion of convergence.

2.3.1 E-Step

The E-Step of the algorithm estimates the variational variables which represent block memberships Z_{iq} of the nodes i as well as C_q which are analogous to the "memberships of memberships". First we describe the estimation procedure for the variational approximations τ_{iq}, next we describe the estimation of signal-noise differentiation probabilities P_q. This two-step procedure differs from prior work because of an additional hierarchical estimation step of the higher-level variational variables P_q.

To estimate the membership vectors, a iterative fixed-point approach is used to estimate τ_{iq}, wherein the derivative for each τ_{iq} is taken based on model parameters and τ_{jl},

$$\log(\tau_{iq}) \propto \log(\alpha_q) + \sum_{k \leq K} \sum_{j \leq n} \tau_{jl} \Bigg(P_q f(X_{ij}^k, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) + (1 - P_q) f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}) + \sum_{l \leq Q; l \neq q} f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}) \Bigg) - 1 + P_q \log \Psi + (1 - P_q) \log(1 - \Psi).$$

After exponentiating, the fixed-point equation can feasibly be solved after the iterating the system until relative stability. This is the same approach as most existing literature (Daudin et al., 2008; Mariadassou et al., 2010). P_q are calculated as follows:

$$\widehat{P_q} = 1 - \Bigg(1 + \Bigg[\exp\Bigg(\sum_{k\leq K}\sum_{i,j\leq n}\tau_{iq}\tau_{jq} \Bigg(f(X_{ij}^k, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) - f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})) + \log\left(\frac{1-\Psi}{\Psi}\right)\Bigg)\Bigg)\Bigg]^{-1}\Bigg)^{-1}. \tag{2.14}$$

Calculations for each of these terms are provided **below**. This section gives derivations for every step of the Variational EM algorithm. First we describe optimizing membership probabilities in the E-Step. We find optimal values for each τ_{iq} by solving this following equation:

$$\begin{aligned}\frac{\partial}{\partial \tau_{iq}}\mathcal{L} =& \log(\alpha_q) + \sum_{k\leq K}\sum_{j\leq n}\tau_{jl}\Bigg(P_q f(X_{ij}^k, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) + (1-P_q) f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}) \\ &+ \sum_{l\leq Q: l\neq q} f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})\Bigg) - \log(\tau_{iq}) - 1 + P_q \log\Psi + (1-P_q)\log(1-\Psi) \\ :=& 0,\end{aligned}$$

rearranging τ_{iq} we solve this equation using a fixed point iteration procedure.

Following estimation of the membership probabilities, the noise probabilities are also estimated in the E-step. Variational variables P_q that serve as the "soft" versions of C_q can be approximated by estimating the probability of block q being a "signal" block or noise block. The terms $\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\log f(\mathbf{X}|\mathbf{Z}), \mathbb{E}[\log R(\mathbf{C})]$, $\mathbb{E}[\log S(\mathbf{C}|\mathbf{Z})]$ in $\mathcal{L}'$ are dependent on $\mathbf{C}$. Practically, because we need to normalize for N_q, which is $1 - P_q$, that variable is more simple (if not the only possible tractable option).

$$\begin{aligned}\frac{\partial}{\partial N_q}\mathcal{L}' =& \frac{\partial}{\partial N_q}\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}[\log f(\mathbf{X}|\mathbf{Z})] - \log N_q + \log(1-N_q) - (\log\Psi + \log(1-\Psi))\sum_i \tau_{iq} \\ :=& 0\end{aligned}$$

where the first term is $f(\cdot)$ is the portion of the multivariate normal density:

$$\sum_k\sum_{i,j}\tau_{iq}\tau_{jq}\Bigg(f(X_{ij}^k, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) - f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}) + \log\left(\frac{1-\Psi}{\Psi}\right)\Bigg) = \log\left(\frac{N_q}{1-N_q}\right)$$

So then, after rearranging:

$$\widehat{N_q} = \left(1 + \left[\exp\left(\sum_{k,i,j} \tau_{iq}\tau_{jq}\left(f(X_{ij}^k, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) - f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})) + \log\left(\frac{1-\Psi}{\Psi}\right)\right)\right)\right]^{-1}\right)^{-1}.$$

Then the final N_q estimates are made after normalizing all $\widehat{N_q}$ such that they sum to one. Finally, the P_q estimates are made by subtracting N_q from 1.

2.3.2 Stochastic Variational Inference

To speed up computation, we apply stochastic variational inference (SVI) to calculate the membership parameters τ_{iq} and P_q. We subsample nodes at each step of the E-step in variational EM. Calculating $\tau_{iq,t}$ and $P_{q,t}$ comprise two stochastic sub-steps of the E-step at iteration step t; we label their SVI estimates as $\widehat{\tau}_{iq,t}$ and $\widehat{P}_{q,t}$. At each t, we sample a set of nodes $M = \{i_1, ..., i_m\}$ of size m and their associated edges from graph layers $\mathbf{X}^1, ..., \mathbf{X}^K$. Let $\tau_{iq,t}^m$ represent the randomly subsampled graph at iteration step t.

1. (Calculating $\tau_{iq,t}^m$) Partial updating step for $\tau_{iq,t}^*$ at time t wherein the subsampled memberships $i, j \in M$ are found:

$$\tau_{iq,t}^* \propto \exp\Bigg(\log(\alpha_q) + \sum_{k\leq K}\sum_{j,l\in M} \tau_{jl,t-1}\Bigg(P_q f(X_{ij}^k, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) + (1-P_q) f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN}) + \sum_{l:l\neq q} f(X_{ij}^k, \boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})\Bigg) - 1 + P_q \log\Psi + (1-P_q)\log(1-\Psi)\Bigg).$$

   The update step averages the newly calculated $\tau_{iq,t}^*$ with the previous value

$$\widehat{\tau}_{iq,t} = \delta_t \tau_{iq,t}^* + (1-\delta_t)\widehat{\tau}_{iq,t-1}.$$

2. (Calculating $P_{q,t}$) The signal probability P_q is calculated in (2.14) but with the same subsampled replacements as done in the previous calculation of $\boldsymbol{\tau}$. For each time point the new noise probability $p_{q,t}^*$ is calculated and averaged with the previous noise probability at time

$t-1$. The update step is

$$\widehat{P}_{q,t} = \delta_t P^*_{q,t} + (1-\delta_t)\widehat{P}_{q,t-1}.$$

To apply stochastic variational inference, we first define the time-variable step size δ_t to retain some memory from previous iteration. A time-varying $\delta_t \in (0,1)$ is selected to satisfy the convexity assumption of (1) $\sum_t \delta_t = \infty$ and (2) $\sum_t \delta_t^2 < \infty$ as outlined in (Hoffman et al., 2012), for some $\kappa \in (.5, 1)$

$$\delta_t = (t+1)^{-\kappa}.$$

However, this criteria needs to be changed when the stochastically sampled variables represent memberships. Empirically, the samples converge at a fast rate when the initial "burn in" steps are subsampled, with subsample sizes increasing with each successive step. If subsampling does not take place, a potentially major impediment may arise from the slow computation speed in early steps where initialized estimates are not near the optimal values. As such, the step sizes are set as such:

$$\delta_t = \min\left(a + \left(\frac{t}{t+1}\right)^{\kappa} n, n\right).$$

a and κ are constants. a governs the smallest subsample size and $\kappa > 1$ governs the rate of increase for subsample size at each step size, with the maximum possible subsample size n. A larger a means a larger starting subsample, and a larger κ means a faster rate of increase in subsample size.

Empirically, for a wide range of simulations, an effective choice for a is between 100 to 200 (depending on network size) and for κ is 2. These values are chosen to ensure computational efficiency in addition to accuracy: computation times for initial values are much slower if the parameter estimates are far from the optimal values which maximize the ELBO, so smaller sample sizes in earlier iterations will economize computation by producing more local minima, while later iterations will yield more globally accurate estimates (Hoffman et al., 2012).

2.3.3 M-Step

Similar to its estimation estimates in Daudin et al. (Daudin et al., 2008), α_q are estimated as follows using Lagrangian multipliers: $\hat{\alpha}_q = \sum_{i,q} \tau_{iq}/n$. The closed-form estimate for the *local* parameters for the mean vector $\boldsymbol{\mu}_q$ for each block q from the M-step is

$$\widehat{\boldsymbol{\mu}}_q = \frac{\sum_{i,j} \tau_{iq}\tau_{jq}\mathbf{X}_{ij}}{\sum_{i,j} \tau_{iq}\tau_{jq}} P_q + \boldsymbol{\mu}_{AN}(1 - P_q).$$

In the above, and all subsequent expressions in this subsection, the derivations are located in Appendix 2.3.4. Similarly to mean calculations, the variance calculations (along diagonals) are

$$\widehat{\boldsymbol{\Sigma}_q} = \frac{\sum_{i,j} \tau_{iq}\tau_{jq}(\mathbf{X}_{ij} - \boldsymbol{\mu}_q)^2}{\sum_{i,j} \tau_{iq}\tau_{jq}} P_q + \boldsymbol{\Sigma}_{AN}(1 - P_q).$$

The cross-term for two layers h, k is written as:

$$\widehat{\boldsymbol{\Sigma}}_{hk,q} = \frac{\sum_{i,j} \tau_{iq}\tau_{jq}(X^k_{ij} - \boldsymbol{\mu}_{q,k})(X^h_{ij} - \boldsymbol{\mu}_{q,h})}{\sum_{i,j} \tau_{iq}\tau_{jq}} P_q.$$

The element-wise correlations at iteration t across layers h, k ($h \neq k$) are then calculated, and the maximum (if $K > 2$) of these values is taken as the putative correlation (across all layers) for block q

$$\hat{\rho}_q = \max_{h,k} \frac{\widehat{\Sigma^q_{hk}}}{\sqrt{\widehat{\Sigma^h_q}\widehat{\Sigma^k_q}}}.$$

If $K = 2$ then no maximum needs to be taken. This is an operational step of the optimization and does not necessarily yield closed-form estimates. However, we note that this value is identical to the *mutual coherence* of estimated correlation matrix and serves as a summary statistic of the estimates for correlations that is consistent with the approximation of the optimization problem we solve with VEM (Tropp, 2006). Theoretical properties of these relationships should be explored in future work.

To calculate the global parameters, the global noise probability term Ψ defined previously is

$$\widehat{\boldsymbol{\mu}}_{AN} = \Psi \frac{\sum_{j,i} \sum_{l,q:q\neq l} \tau_{iq}\tau_{jl}\mathbf{X}_{ij}}{\sum_{j,i} \sum_{l,q:q\neq l} \tau_{iq}\tau_{jl}} + (1-\Psi)\frac{\sum_{j,i} \sum_{q} \tau_{iq}\tau_{jq}(1-P_q)\mathbf{X}_{ij}}{\sum_{j,i} \sum_{q} \tau_{iq}\tau_{jq}(1-P_q)}. \tag{2.15}$$

The covariance term for global noise, as stated earlier, is zero. The variance of global parameters is similarly calculated as:

$$\widehat{\boldsymbol{\Sigma}}_{AN} = \Psi \frac{\sum_{j,i} \sum_{l,q:q\neq l} \tau_{iq}\tau_{jl}(\mathbf{X}_{ij}-\boldsymbol{\mu}_{AN})^2}{\sum_{j,i} \sum_{l,q:q\neq l} \tau_{iq}\tau_{jl}} + (1-\Psi)\frac{\sum_{j,i} \sum_{q} \tau_{iq}\tau_{jl}(1-P_q)(\mathbf{X}_{ij}-\boldsymbol{\mu}_{AN})^2}{\sum_{j,i} \sum_{q} \tau_{iq}\tau_{jq}(1-P_q)},$$

Derivations for these expressions are in Section 2.3.5.

2.3.4 Derivation of Signal Terms for M-Step

The closed-form estimate of the parameter for the mean vector $\boldsymbol{\mu}_q$ for each block q from the M-step is

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_q &= \frac{\sum_{i,j} \tau_{iq}\tau_{jq}\mathbf{X}_{ij}}{\sum_{i,j} \tau_{iq}\tau_{jq}} P_q + \frac{\sum_{i,j} \tau_{iq}\tau_{jq}\boldsymbol{\mu}_{AN}}{\sum_{i,j} \tau_{iq}\tau_{jq}} \cdot (1-P_q) \\ &= \frac{\sum_{i,j} \tau_{iq}\tau_{jq}\mathbf{X}_{ij}}{\sum_{i,j} \tau_{iq}\tau_{jq}} P_q + \boldsymbol{\mu}_{AN}(1-P_q)\end{aligned}$$

Assuming convergence of P_q to either 0 or 1 within the context of the variational iterations, the theoretical value of

$$\boldsymbol{\mu}_q = \begin{cases} \frac{\sum_{i,j} \tau_{iq}\tau_{jq}\mathbf{X}_{ij}}{\sum_{i,j} \tau_{iq}\tau_{jq}} & \text{if } q \text{ is Signal: } P_q = 1 \\ \boldsymbol{\mu}_{AN} & \text{if } q \text{ is Noise: } P_q = 0 \end{cases}$$

Similarly to mean calculations, the variance calculations (along diagonals) are :

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}_q} &= \frac{\sum_{i,j} \tau_{iq}\tau_{jq}(\mathbf{X}_{ij}-\boldsymbol{\mu}_q)^2}{\sum_{i,j} \tau_{iq}\tau_{jq}} \cdot P_q + \boldsymbol{\Sigma}_{AN} \cdot (1-P_q) \\ &= \begin{cases} \sum_{i,j} \tau_{iq}\tau_{jq}(\mathbf{X}_{ij}-\boldsymbol{\mu}_q)^2 / \sum_{i,j} \tau_{iq}\tau_{jq} & \text{if } q \text{ is Signal: } P_q = 1 \\ \boldsymbol{\Sigma}_{AN} & \text{if } q \text{ is Noise: } P_q = 0 \end{cases}\end{aligned}$$

The cross-term for two layers h, k is written as:

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}_{hk,q}} &= \frac{\sum_{i,j} \tau_{iq}\tau_{jq}(\mathbf{X}_{k,ij} - \boldsymbol{\mu}_{q,k})(\mathbf{X}^h_{ij} - \boldsymbol{\mu}_{q,h})}{\sum_{i,j} \tau_{iq}\tau_{jq}} \cdot P_q + 0 \cdot (1 - P_q) \\ &= \frac{\sum_{i,j} \tau_{iq}\tau_{jq}(\mathbf{X}^k_{ij} - \boldsymbol{\mu}_{q,k})((\mathbf{X}^h_{ij} - \boldsymbol{\mu}_{q,h})}{\sum_{i,j} \tau_{iq}\tau_{jq}} \cdot P_q\end{aligned}$$

The element-wise correlations at iteration t across layers h, k ($h \neq k$) are then calculated as

$$\hat{\rho_q}^{h,k} = \frac{\widehat{\Sigma^q_{hk}}}{\sqrt{\widehat{\Sigma^h_q}\widehat{\Sigma^q_k}}}.$$

Finally, the putative correlation (across all layers) for block q is

$$\hat{\rho_q} = \max_{h,k} \hat{\rho_q}^{h,k}.$$

2.3.5 Derivation for Noise Terms in M-Step

To calculate the global parameters (2.3.3), the global noise probability term Ψ defined previously is

$$\begin{aligned}\widehat{\boldsymbol{\mu}_{AN}} &= \mathbb{E}_{R_{\mathbf{X}}(\mathbf{Z},\mathbf{C})}\big[\boldsymbol{\mu}_{AN}\big] \\ &= \mathbb{P}(B_q \text{not } NB)\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\big[\boldsymbol{\mu}_{AN}\big|B_q \text{ is not } NB\big] \\ &\quad + \mathbb{P}(B_q = NB)\mathbb{E}_{R(\mathbf{Z},\mathbf{C})}\big[\boldsymbol{\mu}_{AN}\big|\ \{B_q = NB\}\big]; \quad q : 1 \leq q \leq Q \\ &= \Psi \frac{\sum_{j,i}\sum_{l,q:q\neq l} \tau_{iq}\tau_{jl}\mathbf{X}_{ij}}{\sum_{j,i}\sum_{l,q:q\neq l} \tau_{iq}\tau_{jl}} + (1 - \Psi)\frac{\sum_{j,i}\sum_q \tau_{iq}\tau_{jq}(1 - P_q)\mathbf{X}_{ij}}{\sum_{j,i}\sum_q \tau_{iq}\tau_{jq}(1 - P_q)},\end{aligned}$$

$\widehat{\boldsymbol{\Sigma}_{AN}}$ can also be calculated in a similar way. We describe the derivation of Ψ that was first defined in Definition 2.4: let $\{NB\}$ represent the event that there exists a Noise Block in the multilayer

graph system. The we write the indicator for this event as $\mathbf{1}(NB)$ with probability $\mathbb{P}(NB)$.

$$\begin{aligned}
\Psi &= \mathbb{P}(B_q \neq NB; \forall q : q \leq Q) \\
&= \mathbb{P}(C_q = 1; \forall q : q \leq Q) \\
&= 1 - \mathbb{P}(\text{Global average rate of } q \text{ s.t. } C_q = 0; \forall q : q \leq Q) \\
&= 1 - 1/Q \\
&= (Q-1)/Q
\end{aligned}$$

2.4 Empirical Performance of Synthetic Experiments

In this section we describe the simulation studies to demonstrate the accuracy and efficacy of the proposed method. We design three different experiments for assessing several criterion to evaluate the efficacy of our model.

1. Experiments on many synthetic networks of differing parameters and block sizes to assess membership and parameter recovery, as well as computation time
2. Experiments on many synthetic networks of the *same* parameters and block sizes to assess parameter recovery
3. Simulate a single multilayer network and run under multiple Q to assess a method based on *Integrated Complete Likelihood* (ICL) to determine the optimal block sizes

We considered networks of two and three-layers with sizes $n = 200$ to 500. The complexity of the estimation algorithm scales non-linearly with nodes and layers, but is more efficient and parsimonious compared to existing methods described in the following Section 2.4.4. Computation time for simulations are feasible in networks of several thousand nodes and is suitable for the primary case study, of which the sample size number around 5000.

First, we simulate many small to medium networks with differing underlying memberships and parameters. We then run the `SBANM` algorithm on these networks to demonstrate that the method is able to recover simulated memberships and parameters. We also assess the computation times of various simulations and compare them to existing methods. Secondly, we generate many synthetic networks with the same parameters and memberships and apply `SBANM` to systematically recover

the parameters under more controlled conditions. Finally, we simulate a single small network and run the algorithm under several different settings for the estimate of blocks Q and validate the model selection procedure.

2.4.1 Experimental Procedure

The goal of these experiments is to demonstrate that the proposed method can faithfully recover generated memberships and parameters in a time-efficient manner. As described above, we use two simulation schemes to evaluate membership and parameter recovery (then perform another experiment to assess optimal number of blocks in Section 2.4.5). In all of the experiments outlined above, blockwise parameters for every network are first randomly generated for every layer, then observations (edges) are simulated from multinormal distributions governed by these parameters. Each network has distribution $N_K(\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})$ governing both noise block NB and interstitial noise IN. `SBANM` is then applied to these networks and membership (as well as parameter) recovery is assessed. Simulations are all drawn from differing parameters to demonstrate that the method is effective for a variety of settings different parameter values.

After the ground-truth parameters are generated, we proceed to the second data-generating step. For each mean-covariance pair corresponding to a block, we generate multivariate Gaussian distributions with a sample size of $n_q(n_q-1)/2$, then we convert these multivariate data to weighted edges. Finally, a sample of the AN distribution with size

$$n_{IN} := (n-1)/2 - \sum_{q:q\leq Q} n_q(n_q-1)/2$$

is generated for all n_{IN} interstitial edges between differing blocks.

In all of the experiments, the algorithm is initialized by applying spectral clustering on the sum graph $\tilde{X}$ across all K layers, such that each entry in a single flattened graph $\tilde{\mathbf{X}}$ is $\tilde{X}_{ij} = \sum_{k\leq K} X^k_{ij}$. Another option is drawing that every τ_{iq} is drawn from a uniform distribution, then normalized. Matias et al. propose averaging the graphs and then running k-means over the averages (Matias and Miele, 2017). We initialize by first averaging the layers to $\tilde{X}_{ij}$, then by using spectral clustering (Rohe et al., 2011), which approximates the community structure in a single network quickly and reliably.

2.4.2 Recovery Under Differing Parameters (First Experiment)

In the first experiment, we fix Gaussian priors and generate different multinormal distributions from these hyperparameters, such that every network has different parameters. We generate bivariate networks of size 500 and trivariate networks of size 200 with block sizes between 3 and 5. Block-memberships are generated from a multinomial distribution.

Synthetic data are generated from a two-step procedure. In the first step, Gaussian parameters are randomly generated using fixed priors. In the second step, multivariate Gaussian distributions are generated from the parameters obtained in the first step. The number of blocks Q is first randomly generated. Means and variances of each block, as well as the global mean and variance for the ambient noise, are then independently generated from normal distribution (ie. Gaussian prior), and a positive correlation coefficient is sampled from a uniform distribution between 0 and 1. The first block of each network is designated as NB and its mean and variance follow those of AN. Group sizes n_q for each block are generated from multinomial distributions that were drawn from Dirichlet priors. In order to induce separability of blocks during simulations, we only select the networks whose blocks' minimum Bhattacharya distances are above a certain threshold.

We denote *exact recovery* as whether the `SBANM` algorithm is able to correctly impute and place all the block memberships of the network that was generated based on the multinormal simulation scheme in Section 2.4.1 (Abbe, 2017). Exact recovery rates of the algorithm (for memberships) were fairly accurate. Results show that bivariate simulations induces nearly a 100% (49/50) recovery rate; and 75% (37/50) for the trivariate simulations. In the triavriate case, the imperfect recoveries do recover *most* of the parameters and memberships as shown by existing metrics for community detection in Table 2.1. We note the sensitivity of the recovery rates to the increase in dimensions (or layers); and hints at some parallels with the curse of dimensionality for community detection in multilayer networks (Ertöz et al., 2003), or perhaps due to small sample size of the networks ($n = 200$). Increasing dimensions tends to induce more probable mixtures between blocks that are close together.

Parameter estimates are also reasonably retrieved from the `SBANM` algorithm, both in absolute and relative terms. Mean errors are centered around zero as to not show any systemic bias; absolute percentage differences between ground truths and their estimates hover around 10-25%; some of

the discrepancies may arise from small ground truth values or imperfectly matching memberships. More details can be found in Figure 2.3.

We describe the simulation scheme of the first experiment. The means for each unique block for every network are randomly generated from a Gaussian distribution centered around 0 and 2 respectively for the first and second layers. After the parameters are generated, the observations are simulated from multinormal distributions governed by these parameters. Each network has AN governing both a single block NB and interstitial noise IN that is centered around (-1,0). We repeat this procedure for trivariate networks of $n = 200$ nodes, wherein the Gaussian priors for each (signal) block have means of -2, 0, and 2 respectively for the first, second, and third layers. In order to ensure the separability of blocks during simulations, we only select the networks whose blocks' minimum Bhattacharya distances are above a certain threshold. We calculate the minimum Bhattacharya distances between blocks across 500 simulated networks, and then select the networks with the largest 10% of the minimum Bhattacharya distances to filter out the networks whose blocks are 'far enough away' from each other; we run 50 instances of the `SBANM` algorithm for both the bivariate ($n = 500$) and the trivariate case ($n = 200$).

Results: Fifty runs of the algorithm were performed for both the bivariate and trivariate networks with differing parameters. 500 networks were generated as described in the previous section, then networks with the highest 10% of the minimum Bhattacharya Distances between clusters' parameters are retained.

Though this experiment is primarily focused on *membership recovery*, parameter estimation remains as a byproduct. Across many simulations with a variety of parameters, there does not seem to be much systemic bias in the estimates as empirical means of differences between estimated and true parameters are centered around 0. Median percentage differences, across all estimated parameters, between the estimates and true values are between 20 to 25% for bivariate, and 10-20% for trivariate networks. Histograms for the mean and variance parameters (each distinct parameter is treated like an observation) show essentially matching distributions between estimates and ground truth parameters for means (2.3).

A slight discrepancy between distributions for variance parameters ($\sigma^2_{q,k}$ for $k = 1, 2, 3$) among trivariate networks. This slight bias may be related again to the curse of dimensionality and, while

does not seem to elicit too severe a problem in the clustering results, may be investigated in future endeavors.

Percentage differences between the estimated and ground-truth parameters also show moderately accurate recovery in both bivariate and trivariate networks. The lowest 25% quartiles for all parameters are between 0 and 3 percent and show that these estimates are very close to the ground truths. Conflated with the relatively higher mean and median differences, the low 1st quartiles show that accuracy for parameter runs seem to occur along a binary: either estimates are very close to their targets, or they are fairly far off. Some of the high percentage differences may arise from small ground-truth values, which are divided to calculate percentage differences. Others may arise from the mismatches in clustering memberships. However, this limitation mostly arises in the trivariate case, as there is a near-perfect recovery rate for the bivariate simulations.

2.4.3 Parameter Recovery Under Same Parameters (Second Experiment)

In the second experiment, 100 three-layer networks were generated from a fixed set of parameters as well as memberships ($n = 300$). This experiment with fixed parameters is performed in addition to the first experiment in order to better assess the accuracy of parameter estimates under more controlled conditions. Results show consistently accurate estimates for most of the mean, variance, and correlation parameters (Figure 2.4). Moreover, all memberships were 100% recovered. True parameters are shown in Figure 2.4, and described in more detail in Appendix **??**. The variances for most of the estimates were within 3-5% of the true values, but the estimated variance for ambient noise $\mathbf{\Sigma}_{AN}$ appears to be biased. These types of biases are typical of variational approaches and could be a weakness in VEM for estimating covariance matrices (Mariadassou et al., 2010). Further investigation of this discrepancy may be pursued in future work.

The first experiment was conducted primarily to demonstrated membership recovery under a variety of different parameters and block sizes. The purpose of the second experiment, which runs the algorithm under a set of fixed parameters, is to show that the method recovers parameters effectively. The fixed parameters were generated through simulation with fixed Gaussian distributions with prior means 10,15, and 20 and prior variance parameters of 5. The first entries of each layer correspond to the noise block with fixed means at 5, 10 and 15. The means are: $\boldsymbol{\mu}_{X,q} = (5, 11.98, 11.55, 10.39)$, $\boldsymbol{\mu}_{Y,q} = (10, 16.86, 16.49, 14.81)$, $\boldsymbol{\mu}_{Z,q} = (15, 16.69, 21.25, 21.08)$. The

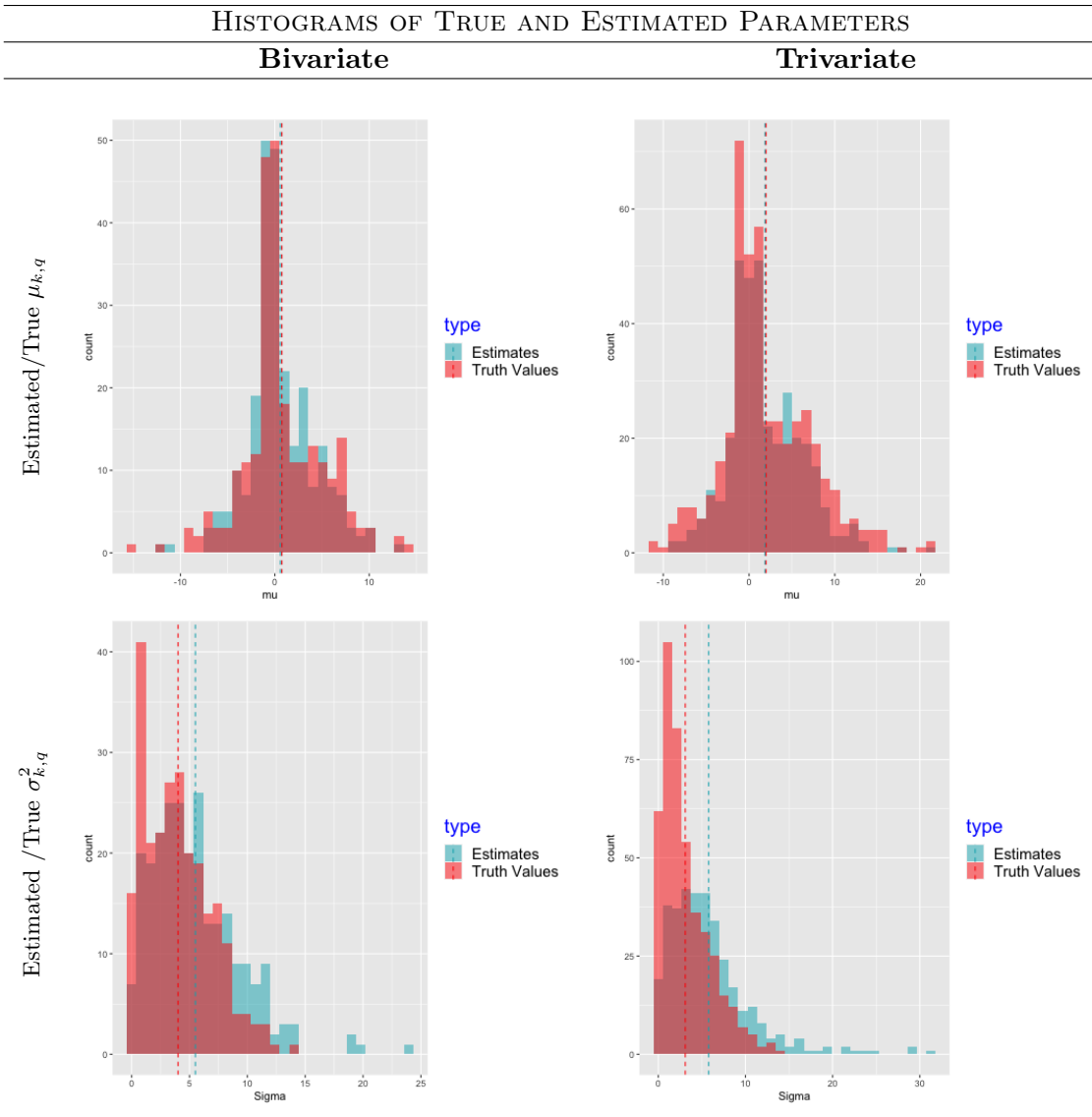


Figure 2.3: Histograms of ground truth (red) and estimate (blue) parameter values for the 2-layer and 3-layer networks compared to the estimated parameters from the algorithm. Parameters across layers are all plotted together. Dashed lines demarcate the empirical means of these estimated and ground truth parameters. For ground truths (red), these empirical means are .75 for $\mu_{k,q}$ (bivariate, top left), 1.98 for $\mu_{k,q}$ (trivariate, top right), 4.01 for $\sigma^2_{k,q}$ (bivariate, bottom left), 3.10 for $\sigma^2_{k,q}$ (trivariate, bottom right). For estimates of parameters, they are .58 for $\mu_{k,q}$ (bivariate, top left), 1.84 for $\mu_{k,q}$ (trivariate, top right), 5.51 for $\sigma^2_{k,q}$ (bivariate, bottom left), 5.58 for $\sigma^2_{k,q}$ (trivariate, bottom right).

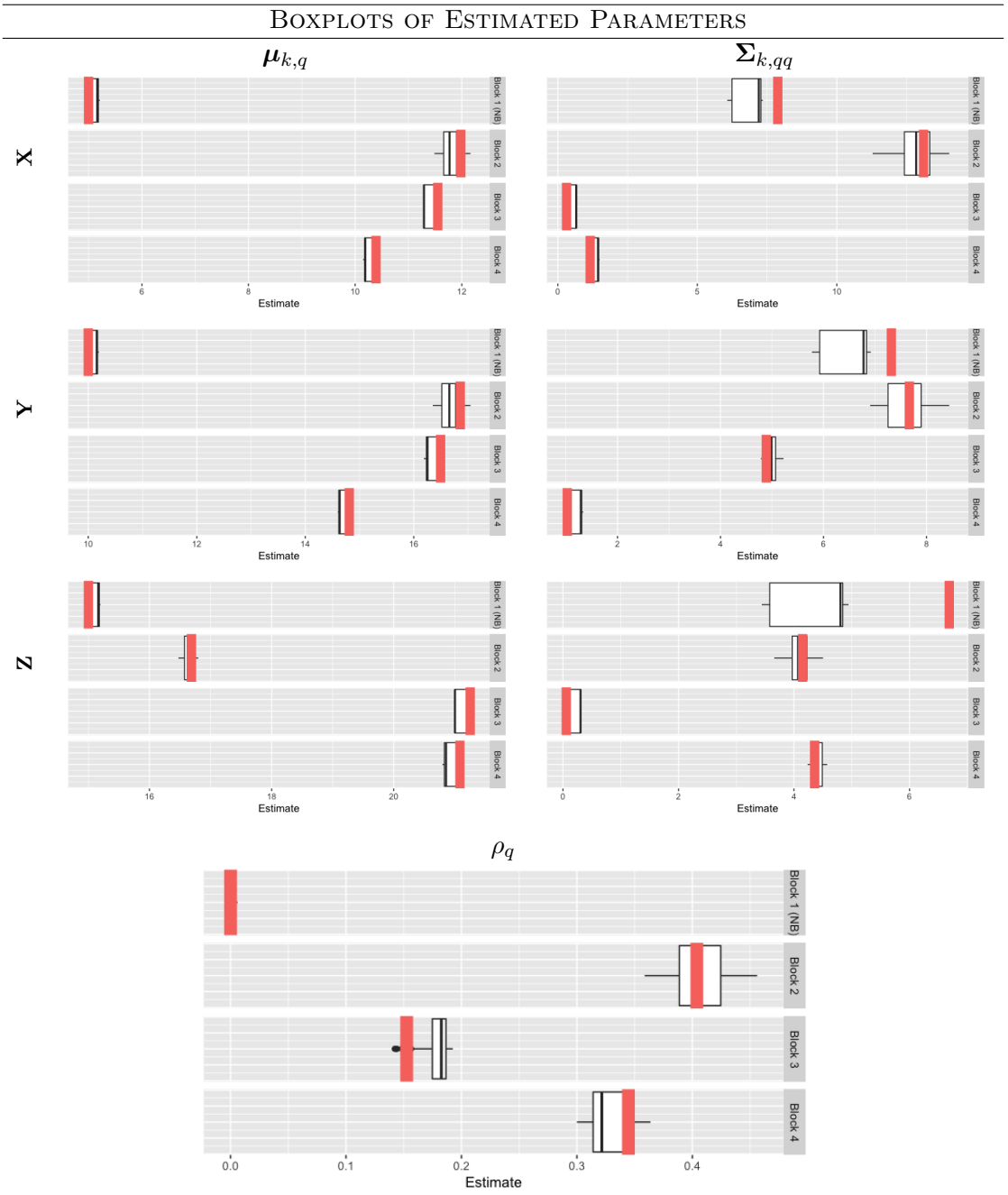


Figure 2.4: Boxplots for repeated estimates of simulations (second type). We ran the algorithm applied to 100 randomly generated networks with the same ground truth parameters and fixed sample sizes. Each boxplot represents the summary of 100 individual estimates corresponding to 100 runs. The red bands represent the ground truth parameters for measn, variances, and correlations.

variances are $\mathbf{\Sigma}_{X,q}$ = (7.88, 13.11, 0.31, 1.16), $\mathbf{\Sigma}_{Y,q}$= (7.32, 7.67, 4.89, 1.03), $\mathbf{\Sigma}_{Z,q}$ =(6.69, 4.15, 0.06, 4.36). The correlations are ρ_q = (0.00, 0.40, 0.15, 0.34), and the true group sizes are 76 nodes for the first block (NB), 97 for the second, 93 for the third, and 34 for the fourth.

Results: We generated 100 networks following these exact specifications and ran `SBANM` on all of them. In Figure 2.4 in the main text, each boxplot comprises a set of 100 estimates for each parameter values. The first row shows those for the first layer (written as $\mathbf{X}$), the second $\mathbf{Y}$, the third $\mathbf{Z}$, and the fourth for correlations between the three layers. The red band shows the true parameter values as listed above.

2.4.4 Comparison with Other Methods

We compared the proposed `SBANM` method with spectral clustering as well as the `dynsbm` proposed by Matias et al. (Matias and Miele, 2017) using the results of the first experiment (Section 2.4.2). We applied spectral clustering 'naively' as in the initialization scheme where all layers are summed and collapsed into a single network because this is an intuitive simple and fast method for multilayer community detection. When we compare to the method with `dynsbm`, we assume two interpretations of their clustering results. Because `dynsbm` imputes different block memberships for every layer, we convert these into cross-layer persistent community labels by (1) taking the most frequent occurrence of the clustered membership across all layers and (2) treating each block-combination across layers as a unique configuration for the definition of a new block. This need to interpret the results of `dynsbm` already reveals an implicit advantage of the `SBANM` method in its inherent parsimony of clusters and interpretability of blocks across layers for certain fitting data-types and scientific questions.

We evaluated *ARI* (Adjusted Rand Index) and *NMI* (Normalized Mutual Information) scores (Wilson et al., 2014; Palowitch et al., 2018; Matias and Miele, 2017). for the three methods with the 50 simulations for both bivariate and trivariate networks and have found that `SBANM` outperforms competing methods in every setting. In the bivariate case, because nearly all simulations yielded *perfect recovery*, the NMI and ARI are both very close to 1. In the trivariate case, the high NMIs and ARIs suggest effective *partial* recovery of the memberships if some of the network block structures are not perfectly recovered. We note that none of the competing methods perfectly recover the block structures for the multigraph systems. We also note that spectral clustering in the bivariate

case outperforms `dynsbm`, but not in the trivariate case; suggesting a potential sensitivity of spectral clustering to the curse of dimensionality.

Method Comparison								
	Bivariate (50 Runs)				**Trivariate (50 Runs)**			
Method	*NMI*		*ARI*		*NMI*		*ARI*	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
`SBANM`	1.00	0.02	1.00	0.01	0.87	0.26	0.87	0.27
Spectral	0.80	0.27	0.84	0.24	0.65	0.31	0.69	0.29
`dynsbm` (unique config.)	0.62	0.25	0.67	0.25	0.75	0.21	0.80	0.21
`dynsbm` (most freq.)	0.68	0.25	0.71	0.25	0.70	0.16	0.77	0.18

Table 2.1: Comparison of different methods for membership recovery using the ARI and NMI measures. `dynsbm` (unique config.) refers to the interpretation of the method when every unique configuration of blocks across layers are treated as a unique block. `dynsbm` (most freq.) treats the block with the most frequent occurence of memberships across all layers as the cross-layer block.

Computing times were higher in `dynsbm` compared to `SBANM` (for spectral clustering, computing time is nearly instant) in both bivariate and trivariate cases. The mean time for trivariate cases is 144 (SD 548) seconds, compared to 160 (125) on average for `dynsbm`. Though `SBANM` computing times have fairly high variance, it is comparable in time to that of `dynsbm` in the trivariate cases. The time differential is much larger in larger bivariate networks. The mean time was 330 (328) seconds for `SBANM` and on average 859 (88) seconds for a few samples of `dynsbm`. The time difference in computation suggests that `SBANM` may better handle larger-size graphs than existing methods. Fitting larger networks when $n > 5000$ are feasible for `SBANM`, but not for `dynsbm`.

2.4.5 Choice of Number of Blocks (Third Experiment)

Model selection in the SBM clustering context usually refers to selection of the number of a priori blocks before VEM estimation as it is the only 'free' parameter in the specification step of the algorithm. Existing approaches (Daudin et al., 2008; Mariadassou et al., 2010; Matias and Miele, 2017) consider the *integrated complete likelihood* (ICL) for assessing block model clustering performance. For this experiment we fix n at the ground-truth Q and apply the method for a range of $\widehat{Q}$ (as the *estimate* for number of blocks). Simulation results show that the usage of ICLs caps at the correct ground truth value and verify that this metric is suitable for evaluation of the method (Figure 2.5.

Model selection in the SBM clustering context usually refers to selection of the number of a priori blocks before VEM estimation as it is the only 'free' parameter in the specification step of

the algorithm. Existing approaches (Daudin et al., 2008; Mariadassou et al., 2010; Matias and Miele, 2017) consider the *integrated complete likelihood* (ICL) for assessing block model clustering performance. Matias et al. write the ICL for multilayer graphs in the following way (adapted to match the notation of this study)

$$ICL(Q) = \log f(\mathbf{X}, \mathbf{Z}) - \frac{1}{2} Q(Q-1) \log(n(K-1)) - pen(n, K, \boldsymbol{\Theta}) \tag{2.16}$$

to translate the terminology, $\boldsymbol{\Theta}$ corresponds to the total set of transition parameters in the SBM, where $\boldsymbol{\Theta} := \boldsymbol{\Theta}_{\text{Signal}} \bigcup \boldsymbol{\Theta}_{\text{Noise}}$ (Matias and Miele, 2017). The penalty parameter $pen(\cdot)$ is chosen dependent on the distributions of the networks; the 'Gaussian homoscedastic' case in Matias et al. is derived to be

$$pen(n, K, \boldsymbol{\Theta}) = Q \cdot \log\left(\frac{n(n-1)K}{2}\right) + \frac{Q(Q-1)}{2} K \cdot \log\left(\frac{n(n-1)}{2}\right).$$

Though the authors made the assumptions that the variances are constant for all blocks, we assume that the models are similar enough to `SBANM` such that the evaluation criterion is applicable to our case. For this portion of the simulation experiment we fix n at 200 and the ground-truth Q at 5. However, we apply the method for a range of hypothesized block numbers $\widehat{Q}$ (as the *estimate* for number of blocks) from 2 to 7. Simulation results show that the usage of ICLs caps at $\widehat{Q} = 5$, the correct ground truth value (Figure 2.5).

Results: We used a single instance of a trivariate network with 200 nodes from the simulations generated in the first experiment. ICLs for five runs of the algorithm were calculated. Each run presupposed a different selection of Q from 2 to 7. The ground-truth value of Q is 5 and Figure 2.5 showed that the ground-truth Q captured the highest ICL.

For large-network simulations, single instances of networks with $n = 1000$ and 2000 are generated for $Q = 4$ and 5. Results yielded exact recovery for memberships and within 5% errors for parameters.

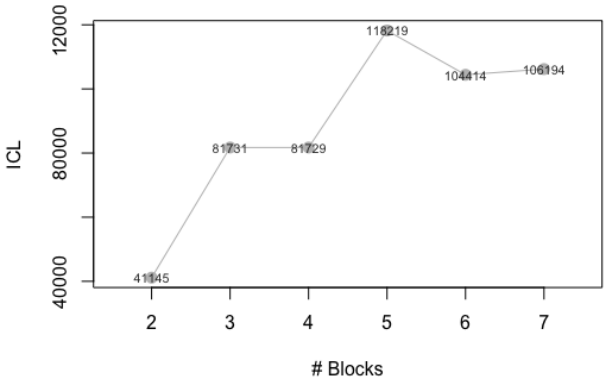


Figure 2.5: ICLs for simulation study for three-layer network of 200 nodes with a ground-truth Q of 5, which maps to the maximum ICL that was found by the method of estimation.

2.5 Case Studies

After validating the method on simulations and real-world datasets, we apply `SBANM` to three different case studies from disparate domains.

2.5.1 Case Study: PNC Psychopathology Networks

We apply `SBANM` to the PNC data which constitutes the primary case study of this chapter. We use networks constructed from *anxiety, behavior*, and *mood* psychopathologies as described in Section 2.1, and then validate the discovered communities from clinical diagnoses for each disorder as well as typical development (TD) and psychosis. We let $\mathbf{X}^x$ represent the layer of symptom response networks for anxiety, $\mathbf{X}^y$ for behavior, and $\mathbf{X}^z$ for mood disorders. Correspondingly, we let $\left(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\mu}_z\right)_{q:q\leq Q}$ represent the means of the edge-connections for each block representing anxiety, behavior, and mood with corresponding standard deviations $\left(\boldsymbol{\sigma}_x, \boldsymbol{\sigma}_y, \boldsymbol{\sigma}_z\right)_{q:q\leq Q}$.

Not much prior work has approached the study of psychiatric networks by constructing networks of individuals as nodes and their similarity as edges. The goal of introducing *ambient noise* to psychopathology symptom networks is to identify groups of people who have similar clinical characteristics and facilitating early identification of individuals who could be at high risk. Existing classification studies on psychosis typically require input from ("training on") already-diagnosed subjects, or psychosis specific symptoms. These methods usually use methods such as logistic regression (Cannon et al., 2016). However, we aim to classify anxiety, mood, and behavior symptoms to identify who is at risk for psychosis *without* the use of psychosis labels in a sample of youth aged

8-21 years, a developmental period prior to the onset of psychotic disorders. Unsupervised analysis is clinically useful in early identification.

We ran the method on youth and early adult data under several different specifications for range of Q. We applied the method to multilayer networks constructed from 5136 youth and 1863 early adult subjects. In each of these runs the `SBANM` algorithm has separated the population into distinct groups with varying block sizes. Table 2.2 shows that highly correlated blocks and NB are discovered with mostly ample separation in terms of Bhattacharya distances as well as post-hoc significance testing (Table 2.6 in Section 2.1.2).

				Parameter Estimates									
	Block	n	ρ_q	μ_x	σ_x	μ_y	σ_y	μ_z	σ_z	$d(N)$	$d(S_1)$	$d(S_2)$	$d(S_3)$
	Youth: 3 Groups												
●	NB	408	0	-0.8	0.3	-0.5	0.4	-0.5	0.3	0.0	3.2	0.5	
●	S_1	2552	0.30	-1.1	0.3	-1.0	0.5	-1.4	0.2	3.2	0.0	1.7	
●	S_2	2176	0.51	-0.6	0.3	-0.5	0.4	-0.0	0.3	0.5	1.7	0.0	
	Early Adult: 4 Groups												
●	NB	48	0	-0.2	0.1	-0.1	0.2	-0.2	0.7	0.0	2.3	0.4	2.2
●	S_1	1495	0.49	-0.9	0.3	-0.7	0.5	-0.6	0.5	2.3	0.0	0.7	2.3
●	S_2	39	0.56	-0.2	0.1	0.1	0.1	-0.4	0.5	0.4	0.7	0.0	1.4
●	S_3	281	0.64	-0.1	0.1	-0.1	0.1	-0.4	0.5	2.2	2.3	1.4	0.0

Table 2.2: Estimated parameters between blocks in youth and early adult subjects, as well as Bhattacharya distances between the blocks. Mean rates for anxiety response networks are represented by μ_x, behavior μ_y, and mood μ_z. Associated standard deviations are also shown.

We used the ICL procedure outlined in Section 2.4.5 to select optimal Q. For youth, the ICL highest for the results when $Q = 3$; for the early adults, $Q = 4$. In both youth and early adults, ICLs suggest that the more parsimonious selections are preferable. In the remainder of this section we mostly focus on the results of these selections of Q, unless there are results specific to the suboptimal-Q model. However, we also note results across model specifications: for example, in youth the same 2552-member cluster is persistent in both settings for Q (3 and 4) (Table 2.2). These results show the persistence of the constellation of symptom agreements across mood, behavior, and anxiety layers.

In general, these results demonstrate the ability of `SBANM` to integrate *anxiety, mood,* and *behavior* symptoms to differentiate groups that signal differential behaviors. Table 2.3 shows the average proportions of subjects who met the criteria of positive symptoms for clinical diagnoses of the anxiety, mood, and behavior disorders as well as psychosis and typically development (TD).

The leftmost columns after block labels and sizes are positive indicators for anxiety, behavior, and mood disorders. They are distinct from symptom data in that each indicator is a binary 'yes' or 'no' for each subject and identified clinically. In nearly all the clustering results, the rates of psychosis spectrum is clearly differentiated among differing clusters. Among youth subjects, S_1 correspond to a group that has relatively low incidence of psychosis (13%). However, NB and S_2 in youth (Table 2.3, left) exhibit similar rates of psychosis spectrum and TD, but with differing anxiety, mood, and behavior symptoms. A table with more selections of Q is found in Table 2.5 in Appendix 2.5.2.

In youth subjects, the S_1 group (in yellow) appears to be have the highest rates of typically developing (TD) youth (Table 2.3) and can be interpreted as a relatively *normal* group. Because it models all between-block interactions, NB may be interpreted as a group that straddles those who exhibit psychosis spectrum symptoms and those who do not. Because this sample that is part symptomatic and part "control" with absence of symptoms, NB may be interpreted a number of different ways in its contrast with correlated signal blocks. Uncorrelated symptoms across all layers potentially signal groups that tend towards psychosis through more individuated channels in NB. In early adult subjects, NB maps to the group with the highest rates of psychosis, as well as the lowest rates of TD subjects.

Psychopathological Symptom Groupings

	Block	n	Anx	Beh	Mood	TD	Psy
	Youth: 3 Groups						
●	NB	408	52	71	14	10	36
●	S_1	2552	37	30	1	44	13
●	S_2	2176	64	55	27	13	44

	Block	n	Anx	Beh	Mood	TD	Psy
	Early Adult: 4 Groups						
●	NB	48	56	52	40	23	56
●	S_1	1495	59	28	23	28	19
●	S_2	39	31	33	5	44	31
●	S_3	281	25	10	7	60	21

Table 2.3: Mean summary statistics for psychiatric diagnoses for youth (left) and early adult (right). The following columns details symptoms of anxiety, behavior, and mood disorders. The 'Psy' column gives the average of whether the respondants have overall diagnoses for psychosis.

The results of clustering demonstrates the ability of `SBANM` to integrate *anxiety, mood,* and *behavior* symptoms to differentiate groups that signal differential, multimodal behaviors. In the results, psychosis rates are clearly differentiated and those in NB are consistently higher. The differential clustering results for youth hints at latent neurodevelopmental pathways for onset of psychosis. Onset of psychosis is characterized by presence of active psychotic symptoms and occurs during early adulthood. It is also better understood as a continuum with patients reporting propor-

tionally more depression, anxiety, and behavior disorders symptoms prior to the onset of psychosis (Cupo et al., 2021). As symptoms segregate with growth and development psychopathology symptom relationships become statistically independent. Clustered subjects with higher correlations ρ_q correspond to the pre-psychotic states of more interconnected pathways, while subjects with independent psychopathologies exhibit more sublimation of psychosis. That these categories emerged without any supervision demonstrates the efficacy of the method to discern risk of developing psychosis. Results also did not show any strong differentiation in demographic characteristics (Table 2.4 in Section 2.5.2).

2.5.2 Additional Posthoc PNC Analyses

Hypothesis tests between different imputed blocks in PNC psychopathological networks (post-processed) and diagnostic categories showed significant differences between all the different clusters. In EA, though the diagnostic comparisons (right) are not all significantly different from each other, the signal (correlated) blocks are all signififcantly different from the noise block NB at the significance level of 0.05.

Demographic characteristics of the clustered subjects are shown in Table 2.4. Rates of patients who are African American, Hispanic, or female are roughly even across the board for most clusters for both youth and early adult under different Q specifications. Regression Z-scores (with respect to psychosis) of demographic factors do not appear to be significant for any cluster.

In EA subjects, NB appears to have higher rates of psychosis on average. When Q is 4, NB actually maps to the group with the highest rates of psychosis, as well as the lowest rates of TD subjects. When $Q = 5$, however, S_3 appears to map to a more typical group (with 50% TD and 7% psychosis). This cluster (for early adults) mirrors the S_1 group found in youth results; and does not seem to appear when Q is set to 4. In youth subjetcs, S_1 has the highest rates of TD. This observation holds for both 3 and 4 groups, as the groups are identical (Table 2.5 in Section 2.5.2), further demonstrating that the clusters are consistent across different Q.

2.5.3 Analysis of US Congressional Voting

The focus of the study is on the PNC data. However, we also show the model's generality by applying the method to political and human mobility data. We use `SBANM` to find latent patterns

	Block	n	Age	Env	%AA	%L	%F
	MP: 4 Gps						
●	N	247	15	-1	33	6	55
●	S_1	2552	14	14	27	5	49
●	S_2	852	15	-14	41	6	51
●	S_3	1485	15	-2	34	7	57
	MP: 3 Gps						
●	N	408	15	-17	41	6	45
●	S_1	2552	14	14	27	5	49
●	S_2	2176	15	-4	35	7	57
	AP: 5 Gps						
●	N	128	19	-10	35	7	53
●	S_1	2	19	54	0	0	50
●	S_2	338	19	-21	39	7	59
●	S_3	792	19	-3	34	6	59
●	S_4	603	19	-9	37	8	61
	AP: 4 Gps						
●	N	48	19	-28	38	10	52
●	S_1	1495	19	-6	35	7	59
●	S_2	39	19	-41	46	8	56
●	S_3	281	20	-13	37	5	65

Table 2.4: Demographic Characteristics of PNC Results. The columns represent respectively: age, environmental factors (Z-scores multiplied by 100), % African American, % Hispanic (Latinx), and % Female.

	Psychopathology Symptoms						
	Block	n	Anx	Beh	Mood	TD	Psy
	Youth: 3 Groups						
●	NB	408	52	71	14	10	36
●	S_1	2552	37	30	1	44	13
●	S_2	2176	64	55	27	13	44
	Youth: 4 Groups						
●	NB	247	56	50	30	7	44
●	S_1	2552	37	30	1	44	13
●	S_2	852	61	72	18	9	39
●	S_3	1485	64	51	29	15	44

	Psychopathology Symptoms						
	Block	n	Anx	Beh	Mood	TD	Psy
	Early Adult: 4 Groups						
●	NB	48	56	52	40	23	56
●	S_1	1495	59	28	23	28	19
●	S_2	39	31	33	5	44	31
●	S_3	281	25	10	7	60	21
	Early Adult: 5 Groups						
●	NB	128	61	44	45	9	25
●	S_1	2	0	0	100	0	0
●	S_2	338	51	28	16	33	29
●	S_3	792	41	15	1	50	7
●	S_4	603	69	37	44	15	33

Table 2.5: (Full) Mean summary statistics for psychiatric diagnoses. The following columns details symptoms of anxiety, behavior, and mood disorders. The 'Psy' column gives the average of whether the respondants have overall diagnoses for psychosis.

Edge Comparison for Youth (3 Gps)			
B_q Comp.	$\mathbf{X}^x$	$\mathbf{X}^y$	$\mathbf{X}^z$
NB- S_1	0.00(**)	0.00(**)	0.00(**)
S_1-S_2	0.00(**)	0.00(**)	0.00(**)
NB-S_2	0.00(**)	0.00(**)	0.00(**)
Edge Comparison for EA (4 Gps)			
B_q Comp.	$\mathbf{X}^x$	$\mathbf{X}^y$	$\mathbf{X}^z$
NB-S_1	0.00(**)	0.00(**)	0.00(**)
NB-S_2	0.00(**)	0.00(**)	0.00(**)
NB-S_3	0.00(**)	0.00(**)	0.00(**)
S_1-S_2	0.00(**)	0.00(**)	0.00(**)
S_1-S_3	0.00(**)	0.00(**)	0.00(**)
S_2-S_3	0.00(**)	0.00(**)	0.25

Diagnosis Comparison for Youth (3 Gps)				
%Anx	%Beh	%Mood	%TD	%Psy
0.00(**)	0.00(**)	0.00(**)	0.00(**)	0.00(**)
0.00(**)	0.00(**)	0.00(**)	0.09	0.00(**)
0.00(**)	0.00(**)	0.00(**)	0.00(**)	0.00(**)
Diagnosis Comparison for EA (4 Gps)				
%Anx	%Beh	%Mood	%TD	%Psy
0.85	6e-4(**)	0.01	0.57	0.00(**)
0.03	0.12	5e-4(**)	0.07	0.03
0.00(**)	0.00(**)	0.00(**)	0.00 (**)	0.00(**)
9e-4(**)	0.59	0.02	0.05	0.11
0.00(**)	0.00(**)	0.00(**)	0.00(**)	0.60
0.59	2e-4(**)	0.84	0.07	0.22

Table 2.6: Hypothesis tests for the clustered blocks in *Youth* subjects along two different criteria. In the first assessment (left), edges in the weighted network for each layer are treated as a i.i.d sample and compared to other edges using t-tests. In the second assessment, proportions of positive clinical diagnoses are tested across different imputed blocks. Let $\mathbf{X}^x$ represent the network of symptom response similarities for anxiety, $\mathbf{X}^y$ for behavior, and $\mathbf{X}^z$ for mood disorders.

in longitudinal US congressional co-voting data to analyze the static as well as dynamic patterns in co-voting amongst US congressional districts, historically a fruitful domain of network analysis (Cho et al., 2011). We also find clusters in longitudinal aggregations of bikeshare networks, whose stations are represented by nodes. Analysis of zones amongst urban mobility services is elucidating for discovering latent patterns within human geography and demographic trends (He et al., 2020a; Carlen et al., 2019; Cazabet et al., 2017a).

In the voteview data, each layer represents interactions among each congressional session. $(\mathbf{X}, \mathbf{Y})$ represents the 100th and 115th sessions of congress, respectively. n represents the number of congressional seats that are common to all three sessions (new or relabeled seats that were added since the first session are not included) Only two layers are used for this application of `SBANM` to the Divvy data, and $(\mathbf{X}, \mathbf{Y})$ in this case represents the normalized, aggregated trips between 2014-2016 and 2016-2018 respectively. The sample size $n = 547$ describes the total number of stations and each edge weight represents aggregate trips between stations.

We use congressional voting records from *Voteview* to uncover patterns in US congressional voting patterns that may yield more nuanced political groups than party labels (i.e. Democrat, Republican) over time. We use a similar pre-processing step as done for the PNC data to assign measures for co-voting similarities between seats in the US House of Representatives during the 100th, and 115th sessions. Voting similarities between representatives in Congress are represented as weighted edges between nodes (representing members). Each layer corresponds to a different

congressional session. We apply the proposed model to data from the *Divvy* bikeshare system in Chicago called to show the ways that demarcating zones of bikeshare trips change across different years. Trip data for Divvy are publicly available on their respective websites (Divvy, 2019).

The overarching motivation for this application is belied by the assumption that political parties change over time and do not necessarily capture the political "tribes" in the US House of Representatives in the past and the present. Prior work use co-voting patterns in the congress and senate in the United States to demonstrate applications of multilayer SBMs by representing district representatives (or senators) as nodes and their covoting similarities as edges (Wilson et al., 2019; Cho et al., 2011). Though most congressional seats have fixed political parties that are representative of their political alignments, parties are assemblages of many constituents with issues that often fragment or congeal (ie polarize) over time. As such, it is useful to trace and segment the groups that either vote with each other persistently, or change drastically following some significant demographic shift. Clustering different political 'tribes' by their similarities in voting is important for studying and forecasting patterns in US politics. It particular, it may be of interest to look for certain "swing" districts that yield more signal for political analysts to study, compared to the ambient levels of connectivity in politically non-contentious districts.

We procure voting data from *Voteview* (Lewis et al., 2020). We use data from all congressional line items from the 100th (1987-89), and 115th (2017-19) sessions, excluding consensus votes where all votes were 'yes' or 'no'. These sessions sample distinct decadal political milieus in the United States across 30 years and serve as snapshots indicating long-term changes in the political inclinations of congressional districts. Though the number of these districts total 435 presently, differing seats often appear and vanish due to redistricting, and we use the seats that were common to both sessions. The resulting network size n is 393.

We use similarity measures similar to that which was applied to PNC survey data for voting records. Between two district seats, which are represented by nodes i and j, the total votes in agreement (both yes or both no) are summed, then subtracted by the total disagreeing votes and divided by the total votes cast. We convert this correlation-like value, which is between -1 and 1, to a statistic that approximates to a normal distribution by applying the same Fisher transformation used in Section 2.1. Like in other studies (Wilson et al., 2019), consensus votes that have either 100% "yes" or 100 % "no" are omitted.

We ran the algorithm over a range of values for estimated block numbers Q, as was done in Section 2.4.1. As the block sizes increase, the ICL also increases, until $Q := 3$ which is where it appears to attain a maximum. We display the clustering results for 3 blocks are shown in Figure

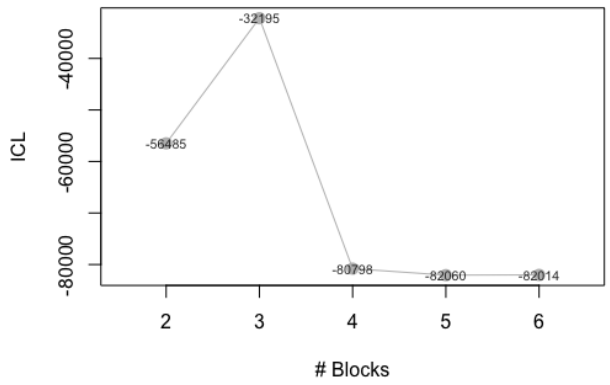


Figure 2.6: Block selection for US congressional voting data based on the method; 3 blocks yields the greatest ICL.

2.6. In addition to the block sizes and estimated correlations, we show the average percentage of Republican party membership ($\%R$) in the 100th and 115th sessions. The results show capture distinct shifts in party membership across the years: NB appears to capture the moderate niche of the congress.

Memberships, Parameters, and Party Affiliation							
Block	n	$\mu_{X,q}$	$\mu_{Y,q}$	ρ_q	*%R(100th)*	*%R(115th)*	Notable People
NB	9	0.02	0.31	0.00	36	67	Nancy Pelosi (1)
S_1	233	0.71	0.36	0.09	4	50	Beto O'Rourke(2), Paul Ryan(2)
S_2	151	0.55	0.45	0.04	99	68	Dick Cheney(1), Liz Cheney(2)

Table 2.7: Clustering results for congressional voting data in the 100th and 115th sessions. In addition to the means and correlations of the (normalized) similarity networks, mean (Republican) party membership rates and notable people in each block are given.

Nine members in NB vote at the same rate with each other as with any other cluster; The interpretation of this block as *moderate* is supported by membership of *moderate Democrat* politicians such as Nancy Pelosi who occupied the seat during empirically verified by the fact that more than half of the block is Republicans in the 115th session. Moreover, NB yields the same rate as every other block votes at the same rate with a different block.

The two biggest political enclaves are large bipartisan *party* that is half Democrat and half Republican in 2015 but was almost entirely Democrat in 1987 (S_1), and another group that was

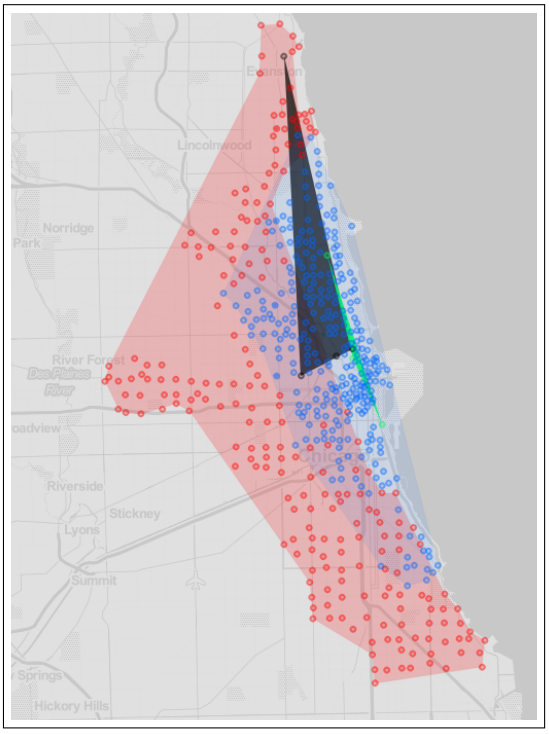
almost entirely Republican in 1987 but only about 2/3 Republican in more recent times. The asymmetry in the blocks S_1 and S_2 is perhaps of note; one can view possibly S_2 as analogous to S_1, but more likely the block is capturing an uneven relationship where there is no Democratic equivalent to the Republican block S_2 which shows entrenchment of voting ideology along geographical (district-wise) lines. These dynamics may be due to fundamental differences in voting patterns between the two parties. Results reveal the large drop-off in the Democrats' political dominance in the 100th session. Instead of capturing static (same-period) blocks, `SBANM` is able to capture some of the largest *differential* movements between the 1980s and 2015.

2.5.4 Human Mobility Data Analysis

The `SBANM` method is applicable to human mobility patterns which is represented by bikeshare data. Bikeshare networks have been argued to trace the latent patterns within human mobility in urban systems (Cazabet et al., 2017a). He et al. (He et al., 2020a) and others have modeled bikeshare stations as nodes and aggregate trips as edges (Carlen et al., 2019), and then gathered conclusions about the patterns of human mobility within these bike-sharing constraints. In particular, prior work have analyzed differences in time-of-day patterns, functional differences (ie work-to-home and home-to-home trips), as well as long-term usage between neighborhoods. Carlen et al. have proposed a time-dependent SBM for (binary) paths between bikeshare stations (Carlen et al., 2019). We convert trip data from the public records of the *Divvy* bikeshare system into time-series networks where each edge represents trips and each node represents stations. We write these network time-series as $\{G_s\}_{1\leq s\leq S}$, where S is the aggregate weekly time-points between January 2014 to June 2016 , and $\{G_t\}_{1\leq t\leq T}$ for T as the aggregate weekly time-points between July 2016 to December 2018, as was done a previous analysis of the *Divvy* system as conducted in He at al. (He et al., 2020b). New stations as well as stations that were removed during this time are omitted, such that the total number of stations ($n = 547$) is consistent across time.

We sum all of the edges across all time points for distinct time-periods S and T The two graphs $\mathbf{X}$ and $\mathbf{Y}$ represent /differential layers across two temporal regimes. We use the number of aggregated trips across each time-regime $\mathbf{X}$ and $\mathbf{Y}$ to represent edge-weights. The edge-weights are then transformed by dividing each value by the respective strengths (sum of weights) to procure a ratio between 0 and 1. The ratio is then converted into an approximately normal value by the *logit*

transformation. Because of this transformation, mean values are negative and between -10 and -20. Estimated statistics (Figure 2.7) are reconverted using the inverse logit transform, then multiplied by the total graphwise sum-of-strengths, to convey a normalized mean rate of trips across stations within the same community.



Parameter Estimates					
		n	μ_X	μ_Y	ρ_q
●	NB	4	1.22	0.37	0
●	S_1	216	8.48	4.78	0.67
●	S_2	3	17.1	0.21	0.00
●	S_3	295	0.29	0.26	0.87

Figure 2.7: Communities found across 2 time-periods in the *Divvy* Bikeshare networks in Chicago, with associated (normalized) estimates for (normalized) mean rates of trips within the cluster in each time period, as well as correlations.

Results show distinct geographical patterns (Figure 2.7). The red cluster is the largest (at 295 nodes) and represents a distinct baseline group for both time periods with activity that persist across time. The high inter-block correlation of .87 in this block suggests persistent trip interactions across time. The blue cluster represents a smaller (216 nodes) but a more *persistent* area of activity: it has higher means for both the first and second layers than that of S_1 for both time-regimes, and also has a high correlation rate. Because this area is closer to more affluent areas around the lake with more parklike amenities (such as the lakefront bike path), this block signifies zones with higher trip activity across both time periods.

Smaller groups NB and S_2 concentrate around the northern part of the city and have very different estimated means that signal drastic change in usage over time. Indeed, the green block

S_2 has the highest first-layer mean μ_X but the lowest second layer mean μ_Y. That the correlation in this block across layers is zero furthermore suggests a disjointingly decreased usage over the two time periods. NB is represented by the grey-black cluster in the northwest part of the city and has the same parameters of ridership as riders traversing across different blocks; which offers an interpretation to the large, but not infeasible, distance between stations (members) in this block. These discovered clusters have interpretable results and suggests tha viability of the method to human mobility data, after the appropriate transformations.

2.6 Discussion

We have introduced a novel method that is motivated by real-world clinical problems and that offers a data-driven approach for grouping subject psychopathologies. This method may predicate deeper understanding or even discovery of psychosis and schizophrenia based on the principles of statistical network theory. We demonstrated the relative efficacy and accuracy of this model compared to existing methods.

Network data in recent years come in more complex forms, which map to the multitude of ways that data relate with one another. They are particularly synchronous with the rise of availability in more different types of data, with even more complex configurations of community structures. Our primary contribution in this research was to introduce the notion of structured noise to weighted SBMs. Other work has explored cases where between-block transitions are all uniquely parameterized (Matias and Miele, 2017), but they do not account for correlations between layers nor do they separate signal from noise. The proposed model is more parsimonious and reveals more interpretable results in clinical and experimental settings. More details on this parsimony can be found in Section 2.6.2. In practice, NB does not represent a control group but rather a dynamic cluster that reflect the noisiest interactions.

2.6.1 Identifiability and Connection to Prior Models

In the introduction, we reference the *affiliation model* in Section 1.2.4 as an example of prior work describing global noise on networks. On a single weighted network, a simple parametric model known as the affiliation model described in Allman et al. (Allman et al., 2011) is formulated as

follows with piecewise global fixed rates:

$$\mu_{ql} = (1 - p_{ql})\delta_0 + p_{ql}F_{ql}\big(\theta_{\text{in}}\mathbf{1}_{q=l} + \theta_{\text{out}}\mathbf{1}_{q\neq l}\big); \quad 1 \leq q, l \leq Q$$

where probability p_{ql} is the sparsity parameter, continuous distribution $F_{ql}(\theta_{ql})$ with parameter θ_{ql} and δ_0 is a dirac mass at zero, and with probability

$$p_{ql} = \alpha\mathbf{1}_{q=l} + \beta\mathbf{1}_{q\neq l}; \quad .$$

One can conceive of the weighted stochastic blockmodel as a special case of the *general form of mixture models for random graphs* described in (Allman et al., 2011). For graph X where each weighted edge is X_{ij} between nodes i, j:

$$\forall q, l \in \{1, ..., Q\} \quad X_{ij}|\{Z_{iq}Z_{jl} = 1\} \sim p_{ql}f(\cdot, \theta_{ql}) + (1 - p_{ql})\delta_0(\cdot),$$

where p_{ql} serves as the sparsity parameter between 0 and 1, which represents the proportion of . $f(\cdot, \theta_{ql})$ represents the parametric family of distributions at specified in group-interactions q and l. The conditional distribution of X_{ij} is a mixture of the Dirac distribution at zero representing non-present edges. The proposed `SBANM` model can also be viewed as an instance of the generalized model above. It is a mixture of the affiliation model and the weighted multilayer SBM. Matias et al. (Matias and Miele, 2017) discuss identifiability of block parameters in multilayer SBMs. The authors cite (Allman et al., 2011) in setting the conditions for identifiability for weighted SBMs over multiple layers. Since the affiliation model is also proven to be identifiable (Allman et al., 2009), we posit that `SBANM` should also be identifiable, but leave more detailed justifictaions in future work.

2.6.2 Parsimony Compared to Other Models

`SBANM` is a parsimonious compared to most other models. If inter-block interactions ($B_q \neq B_l$) are all unique, as in some models (Matias and Miele, 2017; Mariadassou et al., 2010) then this lends to overparametrization, especially at high dimensions ($\approx K \times \frac{Q(Q-1)}{2}$ parameters). The number of parameters may be reasonable for binary and Poisson-distributed multilayer networks, but will quickly inflate in the multivariate Gaussian case. `SBANM` yields $2KQ + Q - 1 + 2K$ parameters

comprising the $2KQ$ mean and (diagonal elements of) variance parameters $\{(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)\}_{q:q\leq Q}$, $Q-1$ correlation parameters $\{\rho_q\}_{q:q\leq Q, q\neq q_{NB}}$, and $2K$ noise parameters $(\boldsymbol{\mu}_{AN}, \boldsymbol{\Sigma}_{AN})$. As Q becomes large, the number of parameters increases quadratically in the canonical weighted SBM but linearly in `SBANM`. As K becomes large, also, the rate of increase for parameters in the proposed method is smaller than that in existing methods. This advantage is demonstrated in computing time comparisons in Section 2.4.4.

2.6.3 Conclusion

We have demonstrated that the method is able to uncover latent, non-trivial patterns in psychiatry (as well as voting and human mobility in Appendices 2.5.3,2.5.4). The application to psychopathology data reflects an ongoing discourse around *nosology* where psychiatric disorders are treated as discrete entities as opposed to multifaceted pathological configurations (van Praag, 2000). Etiologically, the proposed methodology reinforces the multidimensional nature of psychiatric disorders.

Despite its advantages, there remain limitations with `SBANM`. The issue of computation time persistently plagues SBM estimation using VEM. The algorithm slows when K or Q is large. However, in practice it outperforms existing methods. Moreover, usage of stochastic VI has sped up computation time such that previously infeasible sample sizes are made possible. Future work may further explore subsampling methods induce faster computation times.

Ambient noise in networks dovetail the notion of overlapping communities and in particular, SBMs. A class of community detection methods adhere to a *bottom-up* heuristic where sets gradually increase in size until memberships become stable; and naturally allows for separation between *signal* and *noise*. Many of these approaches implicitly assume inherent structure but do not assign an explicitly parametric model to signal or noise (Wilson et al., 2014; Bodwin et al., 2015; Palowitch et al., 2018). Members not assigned to communities are called *background* nodes are identified but not statistically modeled. Uncertainty and ambiguity in block-memberships may be represented by either *noise* or *overlapping blocks*. MMBMs have been useful in modeling real-world data, but in multilayer graphs, overlaps in high dimensions lead to more problems of parameter identifiablity (or altogether avoided (Liu et al., 2018)), and ambient noise serves to assuage the "curse of dimensionality". We refer the reader to the work of Latouche et al. and Airoldi et al. (Latouche

et al., 2011; Airoldi et al., 2007) for background on MMBMs, and leave the connection between *representing noisy signals via overlapping memberships* and global *ambient noise* to future work. Theoretical properties of the model relating to dimensional sensitivities may also be explored.

CHAPTER 3

Community Detection in Weighted Self-Looping Networks

Geographic regions map to social, cultural, and economic structures that enable us to make sense of the world. [1] Demarcation of these regions allows institutional responses to shared problems by creating territorial administrations. These regions are useful at different scales and are created for varying purposes (e.g. cities, places, watersheds, economic regions)(Jones and Paasi, 2013; Paasi, 2013; Pike, 2013). In the United States, metropolitan regions are conceived as collections of counties or equivalent areas (sub-state political units) and are used for different statistical, governance and planning purposes. Yet recent work suggests that these metropolitan regions have coalesced and that *megaregions* spanning multiple states to more effectively project and plan for future growth (Hagler, 2009). At the same time, many urbanized areas are often sub-county regions (Isserman, 2005).

We present a method for inferring geographic regions systematically from the underlying data using community detection methods in network science. One of the key contributions of this approach is to identify multiple overlapping regions at different scales in the same statistical inference framework. We also extend the notion of community to identify nodal regions where peripheral connections are overwhelmed by connections to the core. We also extend community detection methods to include self-loops that have traditionally been implicit or ignored in other community detection work, but are of great importance in commuting networks. The results of these methods identify unusual regions that neither CBSA nor megaregions identify and allow a more nuanced approach to studying and governing metropolitan areas and labor markets (Wheeler, 2013). In summary, our methods are able to differentiate between various types of communities that we classify into three major types:

[1] This chapter is adapted from a manuscript written in 2020 (He et al., 2020b) that was joint work with Shankar Bhamidi and Nikhil Kaza

1. **Monads:** Nodes preferentially attached to themselves in the sense that the self-looping proportions of these nodes are significantly stronger than the baseline self-looping proportions across the entire network.

2. **Nodal communities:** Peripheral nodes more strongly connected to some core nodes rather than among themselves, after accounting for the baseline self-loops.

3. **Non-nodal communities:** Clusters of nodes that are strongly connected to one another. (See Fig. 3.1).

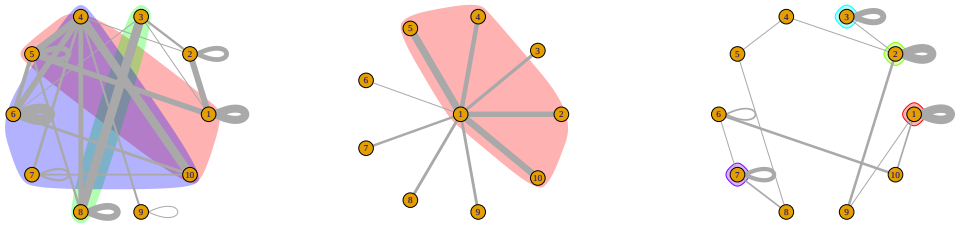


Figure 3.1: Conceptual diagrams representing a) Overlapping non-nodal communities b) Nodal communities (trees) c) monads. The different colors represent different clusters/regions

Networks are used to model the relational structures between individual units of an observed system. A multitude of data structures may be conceived of as networks in the biological, physical, and social sciences. Over the last few years, owing to the explosion in data from a host of areas including social networks, information networks such as the Internet, and biochemical networks such as gene regulatory systems, there has been a concerted inter-disciplinary approach to understanding these data (Newman, 2018a, 2003; Boccaletti et al., 2006; Durrett, 2007; van der Hofstad, 2016). Due to the inherently relational nature of the commuting data, network methods offer a fitting approach for identifying clusters amongst interconnected regions.

3.1 Layout and Contributions

The contributions of this section are twofold: firstly, this section uniquely addresses the *non-standard* format of a network with strong self-looping tendencies. Such a network is specifically tailored to describe the idiosyncracies of commuting behavior between geographical points (U.S.

counties) Secondly, this study is among the first to define single nodes as clusters (*monads*), and also differentiates these clusters from tightly connected communities (cliques) as well as hub-like clusters wherein members are all preferentially connected to a central node, but ncessarily to each others. These categorizations are intricately related to the structure of a strong self-looping network.

Following this section, we first provide some domain background in geographical region demarcation and describe some other approaches for clustering regions. We then describe the data for the specific application of demarcating clusters of counties using US commuting activity in section 3. In section 4, we describe the proposed null model, which accounts for strongly self-looping edge-weights. In section 5, we describe the practical implementation of the community detection algorithm as well as the steps before and after the algorithm implementation such as initialization, differentiation between types of clusters, and filtering of (highly)) overlapping clusters. In section 6, we describe the results of the clustered counties. Finally, in Section 7, we discuss the results of the clustered regions and their implications and relationship to existing delineations.

3.2 Related Work in Region Demarcation

Both CBSA and megaregions implicitly or explicitly rely on the notion of cores. In the case of the former, cores are counties. In the case of the latter, cores are CBSAs. The core-based approach to identifying urban regions has a history dating back to 1950s. Key to this approach is the identification of a core and its connections with the hinterland (nodal communities) (Nystuen and Dacey, 1961). This core-based approach ignores the emerging polycentric structure that has come to dominate regions around the world (Fowler et al., 2018). Since the 1980s, however, further studies have shown that most commuting flows in urban systems are lateral (i.e. between different parts of suburbs and hinterland) rather than core-centered (Plane, 1981). Methods have been proposed to account for these peripheral commuting patterns and used to delineate regions (Tong and Plane, 2014). In our approach, we eschew the a priori identification of cores and instead rely on entire commuting networks, thereby capturing important peripheral connections as well as polynodal and diffuse regions. We identify the core-centered regions in a posthoc analysis.

Non-unique membership is another problem that is not acknowledged in other approaches. Regional delineations tend to partition a set of geographic jfobjects instead of treating them as

members in multiple agglomerations. Especially in densely urbanized regions, many counties have large numbers of commuters to different cities that are relatively close to one another (Han and Goetz, 2019; Kim et al., 2017). Often, delineations (such as OMB) tend to assign counties to one region by breaking membership ties rather than acknowledging connections with multiple regions. It is useful to relax the unique membership condition between an object and the agglomeration to which it belongs.

The other major issue that has received less attention in the literature, both in the context of regional demarcation as well as in the area of network science, is the idea of self-connection. Many agglomeration delineations in both network science and regional science have focused on the connection between two nodes/counties. However, commuting networks have significant self-loops (i.e. commuters within a county): 56% of the total commuters in 2010 in the US commuted within the county where they resided. Ignoring this large commuting pattern skews the results of agglomerations. Since some nodes may be preferentially attached to themselves (measured by the weight of the self-loop), they should be treated as their own agglomerations (see Section 3.5.4).

Community detection methods have been applied to commuting networks to identify regions but traditionally do not account for the above critiques. For example, Nelson and Rae ignore commuting within a node and focus only on commuting between nodes (Nelson and Rae, 2016). They also rely on a community detection algorithm that partitions the entire node set rather than identifying statistically significant connections and clusters. Our proposed approach identifies overlapping communities. Unlike their approach, which starts with census tracts, we start with counties because it is easier to fashion institutions for collections of political boundaries (counties) rather than statistical boundaries (census tracts).

3.3 Data Description

We downloaded our data from the US Census Bureau's Local Origin Destination Employment Statistics (LODES). This dataset contains commuter data between census tracts for all of the continental United States in the year 2010, which we then aggregated to the county level. The data are stored as an undirected and weighted network with self-loops such that each node represents a county and the weight on each edge represents the number of commuters between the connected

counties. Edges with fewer than 100 commuters are removed from the network. Commuters who travel more than 100km (distance between population weighted centroids) are also ignored to remove the effect of telecommuters or super commuters similar to (Nelson and Rae, 2016). The resulting network contains 3,091 nodes and 17,632 edges. Los Angeles County has the largest number of commuters to itself ($\sim$ 3.1 million), while Los Angeles County to Orange County in California is the largest non self-loop edge ($\sim$ 0.6 million). As stated earlier, about 56% of the commuters are commuting within the county. As such, this network can be described as a **strongly self-looping network**.

3.4 Null Model

The Configuration Model, first introduced by Bollobas and Bender (Bollobás, 1980; Bender and Canfield, 1978), is a probability measure on a family of multigraphs that preserves the degree sequence. The input to the model is an observed graph from which we extract the degree sequence, namely a list consisting of the vertices and their corresponding degrees. The model then constructs a *random graph* as follows: start with the degrees of nodes with d_u denoting the degree of vertex u; associate every vertex u with d_u "stubs". One then performs a uniform matching on these stubs to form full edges, thus resulting in a random graph with the prescribed degree sequence but without any other inherent clustering tendency. The relative proclivity of each node to form ties is determined purely upon its degree.

Many of the aforementioned community detection methods utilize the configuration model as the null model (Lancichinetti et al., 2011; Fosdick et al., 2018; Newman, 2006; Girvan and Newman, 2002). We significantly extend the methodology developed by Palowitch et al. (Palowitch et al., 2018) for weighted network data by developing a new null model for **weighted networks with self-loops**. The outcome of the methodology reveals both significantly connected communities, monads, as well as nodal communities in the context of regional commuting flows (see Fig. 3.1).

3.4.1 Notation

We denote an undirected weighted network on n nodes by the triple $G = ([n], \mathbf{A}, \mathbf{W})$, where $[n] := \{1, 2, \ldots, n\}$ is the set of n labeled nodes; $\mathbf{A} = (A_{uv})$ is an $n \times n$ square, symmetric adjacency

matrix with $A_{uv} = 1$ if there is an edge between u and v, and $A_{uv} = 0$ otherwise. Since we are interested in networks with self-loops, we assume $A_{uu} \equiv 1$ for all $u \in [n]$. Though conventionally the self-loop edge is defined as $A_{uu} = 2$, we define it to be 1 as this convention makes the algebra simple when defining the null model. We let $\mathbf{W} = (W_{uv})$ be another symmetric matrix representing (non-negative) weights on edges with W_{uv} denoting the weight between $u, v \in [n]$ with $W_{uv} \equiv 0$ if there is no edge between u and v. We let $d_u = \sum_{v \in [n], v \neq u} A_{uv}$ denote the degree of a vertex, which specifically is the total number of edges connecting to u ignoring self-loops. The **total** strength of a node u is defined as: $s_u = \sum_{v \in [n]} W_{uv}$. We let $d_T = \sum_{u \in [n]} d_u$ and $s_T = \sum_{u \in [n]} s_u$ denote the total degree and weight of the network, respectively. We define ϱ_u to be the propensity of the node to connect to itself by

$$\varrho_u := \frac{W_{uu}}{s_u}, \qquad u \in [n].$$

We define the baseline propensity of the self-loop ratio for the entire network to be:

$$p = \frac{\sum_{u \in [n]} W_{uu}}{s_T}. \tag{3.1}$$

We let $\mathbf{d} = (d_1, \ldots, d_n)$ and $vs. = (s_1, \ldots, s_n)$ denote the degrees and strengths of nodes in $[n]$, respectively.

3.4.2 Continuous Configuration Model Extraction

Significance-based testing was directly pursued in the context of unweighted networks in (Wilson et al., 2014) and weighted networks in (Palowitch et al., 2018). We extend the significance testing based approach developed by Palowitch et al. (Palowitch et al., 2018) in scope and application by adjusting for self-loops. Palowitch et al. (Palowitch et al., 2018) developed a method that used a weighted configuration model as a null model that preserved the expected degrees and strengths of any given node u with its actual respective strengths and degrees. The assumptions of the configuration model are as follows:

$$\mathbf{E}(D(u)) = d_u, \qquad \mathbf{E}(S(u)) = s_u \tag{3.2}$$

We refer to this method as *CCME*.

We start by describing the null model for weighted networks with self-loops that will serve as a comparative model for an observed weighted network $G = ([n], \mathbf{A}, \mathbf{W})$. The model is indexed by a family of parameters $\boldsymbol{\theta} = (\mathbf{d}, vs., \kappa_{SL}, \kappa_{nSL}, a, b)$ where $\mathbf{d}, vs.$ are, as before, the degree and weight sequences of the observed network, respectively, $\kappa_{SL}, \kappa_{nSL} > 0$ are parameters that control the variance of self-loop and non self-loop edge distribution in the null model and $a, b > 0$ are parameters constrained by the relation $a/(a+b) = p$ where p is, as in (3.1), the global self-looping tendency of the observed graph. The concentration parameters a, b of the beta distribution with mean p represent the sparseness and tail shapes (tendencies towards zero or one) of the self-looping probability.

Implicitly, we fix two distributions F_{SL} and F_{nSL} on $\boldsymbol{R}_+$ with **mean one** and variance κ_{SL} and κ_{nSL} respectively. Using the above ingredients we construct a random weighted graph $\mathcal{G} = ([n], \widehat{\mathbf{A}}, \widehat{\mathbf{W}})$ as follows:

(i) **Network topology:** By design $\hat{A}_{uu} = 1$ for all $u \in [n]$. For all $u \neq v$ we let

$$(.\hat{A}_{uv} = 1) = \frac{d_u d_v}{d_T}. \tag{3.3}$$

(ii) **Self-loop edges:** For self-loop edges, we generate edge strengths as follows: First for each vertex $u \in [n]$ (independently across vertices), we generate its *self-loop propensity* $\widehat{\varrho_u} \sim$ Beta(a, b) (i.e. a Beta distribution). Next we generate ξ_{uu} from distribution F_{SL} (independent of $\hat{\varrho}_u$). Then, we model

$$\widehat{W}_{uu} := \widehat{\varrho}_u s_u \xi_{uu}. \tag{3.4}$$

(iii) **Non self-loop edges:** For $u \neq v$ generate edge strengths as follows: first if $\hat{A}_{uv} = 0$ from step (i) then let $\hat{W}_{uv} = 0$. If $\hat{A}_{uv} = 1$ then let $\xi_{uv} \sim F_{nSL}$, and let

$$\widehat{W}_{uv} = (1 - \widehat{\varrho}_u) q_{uv} \xi_{uv}. \tag{3.5}$$

where each q_{uv} represents the following ratio of strengths and degrees of u and v:

$$q_{uv} = \frac{s_u s_v}{s_T} \Big/ \frac{d_u d_v}{d_T}, \tag{3.6}$$

Writing $D(u)$ and $S(u)$ for the degree and strength of vertex u in the associated random graph, it is easy to check that

$$\mathbf{E}(D(u)) = d_u, \qquad \mathbf{E}(S(u)) = s_u, \qquad \mathbf{E}(\widehat{W}_{uu}) = ps_u. \tag{3.7}$$

The weight matrix and adjacency matrix represent inherently different, though correlated, modes of relation. For example, in a social network one can imagine two individuals having similar degrees but very different rates of interaction with the individuals they are connected to. In the context of the commuting data, Mesa County, CO and Los Angeles County, CA have similar degree but very different strengths. Thus part of the aim of this chapter was to develop a baseline null model that would preserve both degrees and strengths as well as a baseline level of self-loopiness and then compare an empirically observed network against this null model to extract regions of significantly higher connectivity after accounting for this baseline connectivity.

We note that in (3.7), the first two conditions are identical to that of the 'ordinary' CCME method, but the third which preserves the ratios of expected self-loops is novel. The model preserves (on average) the degrees and strengths of the observed graph without any other specific notion of clustering. Each vertex has no particular preferential self-looping proclivity other then the average tendency p of the entire network. We refer to this model as CCME with self-loop adjustment (*CCME-SL*).

3.4.3 Parameter Specifications

Palowitch et.al (Palowitch et al., 2018) use a method-of-moments estimator to specify parameters for CCME. We use this method to *learn* the parameters from the observed graph. We specify two types of variables to describe the uncertainty arising out of the strengths of the nodes' connection propensities.

Recall that we denote κ_{SL} as the variance of the self-looping edge weight distribution (with distribution F_{SL}) and κ_{nSL} as the variance of the non-self-looping edge weight distribution (with distribution F_{nSL}). Both of these variables have mean one to ensure the preservation of the strengths and degrees for the configuration null model and to ensure identifiability.

For variance parameters, The method-of-moments estimates for κ_{SL} and κ_{nSL} are as follows:

$$\hat{\kappa}_{SL} = \frac{\sum_{u\in[n]}(W_{uu} - ps_u)^2 - \sum_{u\in[n]} s_u^2\hat{\sigma}_p^2}{\sum_{u\in[n]} s_u^2\left(\hat{\sigma}_p^2 + p^2\right)} \tag{3.8}$$

$$\hat{\kappa}_{nSL} = \frac{\sum_{u\in[n]}\sum_{v\neq u}(W_{uv} - (1-p)q_{uv})^2 - \hat{\sigma}_p^2\sum_{u\in[n]}\sum_{v\neq u} q_{uv}^2}{(\hat{\sigma}_p^2 + (1-p)^2)\sum_{u\in[n]}\sum_{v\neq u} q_{uv}^2}, \tag{3.9}$$

where $\hat{\sigma}_p^2$ represents the estimated variance of ϱ_u using empirical method-of-moments, q_{uv} represents the ratio of strengths to degrees as described in (3.6).

$$\hat{\sigma}_p^2 := \mathrm{Var}(\hat{\varrho}_u) = \frac{1}{n-1}\sum_{u\in[n]}\left(\frac{W_{uu}}{s_u} - p\right)^2. \tag{3.10}$$

$\hat{\kappa}_{nSL}$ represents the variation in relative weights between two edges when we know that the strengths (total sum of weights) are the same. $\hat{\kappa}_{SL}$ represents the variation within a single self-directed edge. These estimates account for the inherent variability of the edge weights of an empirically observed network. The eventual proposed score function (Section 3.5.2) used to judge the significance of the internal connectivity structure of a community (or any subset of nodes) can use this variability metric in its calibration of significance. Details on the derivations of these parameters can be found in the following sections.

3.4.4 Beta Random Variable to Model Self Looping Proportion

We specify ϱ_u as adhering to a Beta(a, b) distribution independent across $u \in [n]$ with mean p and with variance equal to $\hat{\sigma_p}^2$, the sample variance of $\{\varrho_u : u \in [n]\}$. The beta distribution is supported on $[0, 1]$. We designate the proportion of self-looping commuters in each node as comprised of the averages of decisions to either commute *in-county* or *out-of-county* by a host of commuters. The empirical distribution of $\hat{\varrho}_u$ closely matches the simulated values of ϱ_u, except for a few nodes that are at the upper or lower end of the distribution. We note that for a beta distribution,

$$\mathbb{E}(\varrho_u) = \frac{a}{a+b} = p; \quad \mathrm{Var}(\varrho_u) = \frac{ab}{(a+b)^2(a+b+1)} := \sigma_p^2 \tag{3.11}$$

We use p and $\hat{\sigma}_p^2$ (3.10) to determine a and b and, from (3.11), express their estimates as

$$\hat{a} = \frac{-p(p^2 - p + \hat{\sigma}_p^2)}{\hat{\sigma}_p^2}; \quad \hat{b} = \frac{(p-1)(p^2 - p + \hat{\sigma}_p^2)}{\hat{\sigma}_p^2}. \tag{3.12}$$

3.4.5 Variance of ξ_{uu}: κ_{SL}

Note that

$$\begin{aligned}\mathrm{Var}(\xi_{uu}\varrho_u) &= \mathrm{Var}(\xi_{uu})\mathrm{Var}(\varrho_u) + \mathrm{Var}(\xi_{uu})(\mathbb{E}[\varrho_u])^2 + (\mathbb{E}[\xi_{uu}])^2\mathrm{Var}(\varrho_u) \\ &= \kappa_{SL}(\sigma_p^2 + p^2) + \sigma_p^2\end{aligned}$$

as the two variables are assumed to be independent. The sample variance of W_{uu} may be decomposed in the following way:

$$\begin{aligned}\frac{1}{n}\sum_{u\in[n]} \mathrm{Var}(W_{uu}) &= \frac{1}{n}\sum_{u\in[n]} s_u^2 \mathrm{Var}(\xi_{uu}\varrho_u) \\ &= \frac{1}{n}\sum_{u\in[n]} s_u^2 \left(\kappa_{SL}(\sigma_p^2 + p^2) + \sigma_p^2\right) \\ &= \frac{1}{n}\sum_{u\in[n]} s_u^2 \kappa_{SL}\left(\sigma_p^2 + p^2\right) + \frac{1}{n}\sum_{u\in[n]} s_u^2\sigma_p^2\end{aligned}$$

We derive another calculation of the sample standard deviation of self looping weights using a method of moments estimator.

$$\frac{1}{n}\sum_{u\in[n]} \mathrm{Var}(W_{uu}) \approx \frac{1}{n}\sum_{u\in[n]} (W_{uu} - ps_u)^2$$

From the above two equations, we derive the following approximation

$$\frac{1}{n}\sum_{u\in[n]}(W_{uu}-ps_u)^2 \approx \kappa_{SL}\frac{1}{n}\sum_{u\in[n]} s_u^2\left(\sigma_p^2+p^2\right)+\frac{1}{n}\sum_{u\in[n]} s_u^2\sigma_p^2$$

$$\implies \kappa_{SL}\sum_{u\in[n]} s_u^2\left(\sigma_p^2+p^2\right) \approx \sum_{u\in[n]}(W_{uu}-ps_u)^2-\sum_{u\in[n]} s_u^2\sigma_p^2$$

Rearranging the above equation and replacing unknown parameters by their estimates, we derive the estimate for $\hat{\kappa}_{SL}$

$$\hat{\kappa}_{SL}=\frac{\sum_{u\in[n]}(W_{uu}-ps_u)^2-\sum_{u\in[n]} s_u^2\sigma_p^2}{\sum_{u\in[n]} s_u^2\left(\hat{\sigma}_p^2+p^2\right)}$$

3.4.6 Variance of W_{uv}

The properties of the weighted configuration model stipulate that the expectation of an edge weight given that there exist an edge, by equation 3.5, is:

$$\mathbb{E}[W_{uv}|\mathbf{1}(A_{uv})]=(1-p)q_{uv}$$

q_{uv} in the above equation is defined in the main body of the chapter in equation 3.6. We calculate the variance of W_{uv} by decomposing it into two terms **I** and **II** using the following identity

$$\begin{aligned}\mathrm{Var}(W_{uv}) &= \mathbb{E}[\mathrm{Var}(W_{uv}|A_{uv})]+\mathrm{Var}(\mathbb{E}[W_{uv}|A_{uv}])\\ &= \mathbf{I}+\mathbf{II}\end{aligned} \tag{3.13}$$

To calculate **I**, we note that,

$$\mathrm{Var}(W_{uv}|A_{uv})=q_{uv}^2\mathrm{Var}((1-\varrho_u)\xi_{uv}|A_{uv})$$

Therefore, calculating **I** first necessitates calculating the conditional variance of $(1-\varrho_u)\xi_{uv}$, given that there exists an edge, under an expectation. As shorthand, we define the operation

$\mathrm{Var}_A(\cdot) := \mathrm{Var}(\cdot|A_{uv})$ and $\mathbb{E}_A(\cdot) := \mathbb{E}_A[\cdot|A_{uv}]$.

$$\begin{aligned}
\frac{1}{q_{uv}^2}\mathbb{E}[\mathrm{Var}(W_{uv}|A_{uv})\mathbf{1}(A_{uv})] =& \mathbb{E}[\mathrm{Var}((1-\varrho_u)\xi_{uv}|A_{uv})\mathbf{1}(A_{uv})] \\
=& \mathbb{E}[(\mathrm{Var}_A(1-\varrho_u)\mathrm{Var}_A(\xi_{uv}) \\
& + \mathrm{Var}_A(1-\varrho_u)\mathbb{E}_A[\xi_{uv}]^2 + \mathrm{Var}_A(1-\xi_{uv})\mathbb{E}_A[1-\varrho_u]^2)\mathbf{1}(A_{uv})] \\
=& \mathbb{E}[(\mathrm{Var}_A(\varrho_u)\kappa_{nSL} + \mathrm{Var}_A(\varrho_u) + \kappa_{nSL}(1-p)^2)\mathbf{1}(A_{uv})] \\
=& \mathbb{E}[(\sigma_p^2\kappa_{nSL} + \sigma_p^2 + \kappa_{nSL}(1-p)^2)\mathbf{1}(A_{uv})] \\
=& \mathbb{E}[((\sigma_p^2 + (1-p)^2)\kappa_{nSL} + \sigma_p^2)\mathbf{1}(A_{uv})] \qquad (3.14)
\end{aligned}$$

Since relation 3.14 holds for all A_{uv}, it becomes apparent that

$$\mathrm{Var}(W_{uv}|A_{uv}) = q_{uv}^2((\sigma_p^2 + (1-p)^2)\kappa_{nSL} + \sigma_p^2) \qquad (3.15)$$

Using equation 3.15, the calculation of **I** becomes straightforward:

$$\begin{aligned}
\mathbf{I} &= \mathbb{E}[\mathrm{Var}(W_{uv}|A_{uv})] \\
&= q_{uv}^2\mathbb{E}[(\sigma_p^2 + (1-p)^2)\kappa_{nSL} + \sigma_p^2)\mathbf{1}(A_{uv})] \\
&= q_{uv}^2 \cdot (\sigma_p^2 + (1-p)^2)\kappa_{nSL} + \sigma_p^2) \cdot \mathbb{P}(A_{uv}). \qquad (3.16)
\end{aligned}$$

The calculation of **II** , similarly, is as follows:

$$\begin{aligned}
\mathbf{II} &= \mathrm{Var}(\mathbb{E}[W_{uv}|A_{uv}]) \\
&= \mathrm{Var}(q_{uv}(1-p)\mathbf{1}(A_{uv})) \\
&= (1-p)^2q_{uv}^2\mathrm{Var}(\mathbf{1}(A_{uv})) \\
&= (1-p)^2q_{uv}^2\mathbb{P}(A_{uv})(1-\mathbb{P}(A_{uv})). \qquad (3.17)
\end{aligned}$$

Putting the two equations 3.16 and 3.17 together, we are able to solve for the variance of W_{uv} from equation 3.13, and substituting the expression for $\mathbb{P}(A_{uv})$ from equation 3.3,

$$\begin{aligned}
\text{Var}(W_{uv}) &= \mathbb{E}[\text{Var}(W_{uv}|A_{uv})] + \text{Var}(\mathbb{E}[W_{uv}|A_{uv}]) \\
&= q_{uv}^2 \cdot (\sigma_p^2 + (1-p)^2) \cdot \kappa_{nSL} + \sigma_p^2) \cdot \mathbb{P}(A_{uv}) + q_{uv}^2 \mathbb{P}(A_{uv})(1 - \mathbb{P}(A_{uv}))) \\
&= q_{uv}^2 \mathbb{P}(A_{uv})((\sigma_p^2 + (1-p)^2) \cdot \kappa_{nSL} + \sigma_p^2 + 1 - \mathbb{P}(A_{uv})) \\
&= r_{uv} \left((\sigma_p^2 + (1-p)^2) \cdot \kappa_{nSL} + \sigma_p^2 + 1 - \frac{d_u d_v}{d_T} \right)
\end{aligned}$$

where

$$r_{uv} = \frac{(\frac{s_u s_v}{s_T})^2}{\frac{d_u d_v}{d_T}} = q_{uv}^2 \mathbb{P}(A_{uv}). \tag{3.18}$$

3.4.7 Variance of ξ_{uv}: κ_{nSL}

Now we construct a similar method of moments estimator for κ_{nSL} as was done for κ_{SL}. However, we also make use of the expression for the conditional variance of W_{uv} given the existence of an edge in equation 3.15.

$$\begin{aligned}
\sum_{u \in [n]} \sum_{v \neq u} \mathbb{E}[(W_{uv} - \mathbb{E}[W_{uv}])^2 | A_{uv}] &\approx \sum_{u \in [n]} \sum_{v \neq u} \text{Var}(W_{uv}|A_{uv}) \\
&= \sum_{u \in [n]} \sum_{v \neq u} ((\sigma_p^2 + (1-p)^2)\kappa_{nSL} + \sigma_p^2 q_{uv}^2) \\
&= (\sigma_p^2 + (1-p)^2)\kappa_{nSL} \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2 + \sigma_p^2 \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2
\end{aligned}$$

After solving for κ_{nSL} in the above equation and substituting unknown variables with their estimates, then changing the approximation to an equation, we derive the estimate $\hat{\kappa}_{nSL}$ for κ_{nSL}, thus obtaining the estimate as given in equation 3.8:

$$\hat{\kappa}_{nSL} = \frac{\sum_{u \in [n]} \sum_{v \neq u} (W_{uv} - (1-p)q_{uv})^2 - \hat{\sigma_p}^2 \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2}{(\hat{\sigma_p}^2 + (1-p)^2) \sum_{u \in [n]} \sum_{v \neq u} q_{uv}^2}$$

3.4.8 Central Limit Theorem for $S(u, B, \mathcal{G})$ in set B

In this section we detail the calculation of the expectation and variance of the relative strength $S(u, B, \mathcal{G})$ of a given node-set B used in iterative testing, described in the body of the text in Section 3.5.2.

$$\begin{aligned} S(u, B, \mathcal{G}) &= \sum_{v \neq u, v \in B} (1 - \varrho_u) q_{uv} \xi_{uv} \\ &= \sum_{v \neq u, v \in B} (1 - \varrho_u) \frac{\frac{s_u s_v}{s_T}}{\frac{d_u d_v}{d_T}} \xi_{uv} \end{aligned}$$

Taking the expectation of each ξ_{uv} gives the following expression for the strength of the node set

$$\begin{aligned} \mathbb{E}[S(u, B, \mathcal{G})] &= (1 - p) \sum_{v \neq u, v \in B} \frac{d_u d_v}{d_T} \frac{\frac{s_u s_v}{s_T}}{\frac{d_u d_v}{d_T}} \\ &= (1 - p) \sum_{v \neq u, v \in B} \frac{s_u s_v}{s_T} \\ &= s_u \left((1 - p) \sum_{v \neq u, v \in B} \frac{s_v}{s_T} \right) \end{aligned}$$

We have found, in Section 3.4.6, that the variance of a given W_{uv} is expressed as:

$$\mathrm{Var}(W_{uv}) = r_{uv} \left((\sigma_p^2 + (1 - p)^2) \kappa_{nSL} + \sigma_p^2 + 1 - \frac{d_u d_v}{d_T} \right)$$

Adding the variance terms together in set B yields:

$$\begin{aligned} \mathrm{Var}(S(u, B, \mathcal{G})) &= \sum_{u \neq v, u \in B} \mathrm{Var}(W_{uv}) \\ &= \sum_{u \in B} r_{uv} \left((\sigma_p^2 + (1 - p)^2) \kappa_{nSL} + \sigma_p^2 + 1 - \frac{d_u d_v}{d_T} \right) \end{aligned}$$

Then given that B is 'typical' and that d_u and B are sufficiently large, that $S(u, B, \mathcal{G})$ is approximately normal. For $\mu(u, B) = \mathbb{E}[S(u, B, \mathcal{G})]$ and $\sigma(u, B)^2 = \mathrm{Var}(S(u, B, \mathcal{G}))$

$$\frac{S(u, B, \mathcal{G}) - \mu(u, B)}{\sigma(u, B)} \implies \mathcal{N}(0, 1) \tag{3.19}$$

Hence in each step of iterative testing in the update step of CCME, the normal p-value is used to iteratively reject insignificant nodes in a candidate community. Assumptions for and proofs of this Central Limit Theorem can be found in (Palowitch et al., 2018).

3.5 Community Detection Algorithm

The CCME-SL algorithm is split into three general phases: *initialization, update,* and *filtering*. These steps compose the general procedure of iterative testing. Significant communities are groups of nodes with cross-edges that deviate considerably from the expected values under a null model. Significant communities are determined by repeatedly applying an iterative search algorithm that starts with a seed set B_0 and finds all nodes with edges connecting to the seed set. The edge-weights are then summed as a test statistic which is evaluated against the expected values of the sums of the weights in the set under the null model (described in Section 3.4.8) with respect to each node in the starting seed set B_0, imputing a p-value for each node.

Each p-value from the candidate set is rejected if it is significant after being corrected by the Benjamini-Hochberg correction. The nodes with p-values that are *significant* in the present iteration are used as the initial seed sets for the next iteration. The final set B is extracted when the node-set becomes stable: when at some iteration step k, $B_k = B_{k+1}$. Nodes in the final set have a stronger affiliation with each other and have fewer edge connections with all other nodes outside the set.

In this section, we describe each of the phases of CCME-SL in detail. We also describe the hub and monad detection steps as post-community detection phases of the method.

3.5.1 Initialization

We initialize (step 1) sets B_0 by setting counties with high commuting volume as seed nodes (which represent counties). We select nodes that have above 20,000 self-commuters as seed nodes. We select these nodes because they are proxies for relative population centers where commuter traffic radiates outwards to more peripheral connections upon each iteration. We then find all

nodes which are connected to each seed node. The seed node and its connected nodes are used as the initializing sets B_0. The seeds are largely irrelevant to the final outcome: the final outcomes reveal similar outcomes regardless of what the initial nodes selected are, so long as a majority of the high-volume nodes are included (see Fig. 4 in Supporting Information). However, because the initial seed nodes are fixed using the above heuristic, the algorithm converges to the same resulting clusters under the same parameters α and τ.

3.5.2 Update

Stable communities are found using an iterative node-set updating scheme based on the p-value of the connectivity between a single node $u \in [n]$ and a candidate (testing) set $B \in [n]$. We denote $S(u, B, G)$ as the *connectivity* of a single node to the set of nodes which is hypothesized to be a community:

$$S(u, B, G) = \sum_{v \in B} W_{uv}.$$

When the observed value $S(u, B, G)$ significantly exceeds the expected sum of weights under the null model, then there is evidence to support the claim that there is some additional structure undergirding the set of nodes than that which is posited by the null model. The null model attributes connectivity between sets of nodes as dependent only on the strengths and degrees of the aggregations of the nodes themselves.

The p-value representing the significance of a node-set is given by:

$$p(u, B, \mathcal{G}) = \mathbb{P}(S(u, B, G) > S(u, B, \mathcal{G})).$$

In the above equation G is observed but $\mathcal{G}$ is random with a distribution given by the null model $\mathbb{P}$ and with each B representing the candidate set to be tested. When the observed value of $S(u, B, \mathcal{G})$ is much larger than the expected value, expressed as $S(u, B, G)$, the p-value is low. Low p-values are rejected in an iterative fashion so as to allow the formation of node-sets with edges that are consistently significantly connected to each other. We define these sets as communities.

The iterative method is described as follows: for each $u \in [n]$, given a set B (or denoted by B_k at k^{th} iteration), we find all counties that are connected to the present set, then p-values are imputed for members of the candidate set B and repeatedly tested until the set becomes stable upon sequential iterations:

(i) Calculate p-values $\mathbf{p} = p(u, B, \mathcal{G})$. P-values are calculated using a normal approximation for the distribution of $S(u, B, \mathcal{G})$. Details on this part of the procedure are given in Section 3.4.8

(ii) Obtain threshold $\tau(\mathbf{p})$ using a Benjamini-Hochberg multiple testing procedure (Benjamini and Hochberg, 1995). The procedure is used for sets of p-values that are obtained through multiple hypothesis testing. The rejection method ensures that the expected number of falsely rejected hypotheses divided by the total number of rejected hypotheses (false discovery rate, or FDR) has a maximum percentage of α. A false discovery rate threshold α of 0.05 is common in many applications, but for community detection we find empirically that such a threshold should be lower to avoid excess overlaps.

(iii) The next set reached by the iteration is defined as $B' = \{u : p(u, B, \mathcal{G}) \leq \tau(\mathbf{p})\}$

The above steps are iterated with B' replacing B until we reach a fixed point. We set $\alpha = 0.01$ at each step of iterative testing to perform community detection. The threshold can be made higher or lower, and such adjustments do not change the results drastically (see Supporting Information for details), but the threshold of .01 appears to be optimal for maximizing coverage and minimizing overlaps.

3.5.3 Filtering

After obtaining M stable communities C^j, where $j = 1, ..., M$, we remove redundant node-sets that have a high proportion of overlap with other sets. Redundancy in clusters is evaluated using the Jaccard similarity index. A Jaccard similarity index of two sets is defined as the ratio of the size of common elements between the two sets over the total distinct elements, or concisely expressed as $J(A, B) = |A \bigcap B|/|A \bigcup B|$ for two sets A, B (Jaccard, 1901).

We evaluate Jaccard similarities for each pair of found communities C^i, C^j. If the Jaccard index $J(C^i, C^j)$ is above a given pre-set threshold τ, then the clusters are redundant and we select

a preferred cluster by calculating the average weight per connection between nodes. We use a simple formula for a given stable node-set C:

$$K(C) = \frac{\sum_{v \in C, v \neq u} \sum_{u \in C} W_{uv}}{|C|}.$$

$K(C)$ roughly measures the average sum of weights among cross-edges per node within a candidate set C. Given that two sets C^i, C^j have Jaccard overlaps larger than τ, a higher $K(C^i)$ compared to $K(C^j)$ signifies more interconnectivity between nodes in C^i and thus it is kept in the final set of communities while C^j is removed. We set the τ parameter to be 0.80 when implementing the method on commuting networks.

3.5.4 Detection of Monads

One unique feature of geographical commuting networks is that a non-trivial proportion of the total commuting volume is not found in edges across vertices because most residents commutes within their counties. We define the degree of *monadicity* of a given node as the following:

$$I_u := W_{uu} - ps_u.$$

Assuming the observed graph originated from the null model, the variable I_u measures two things. Firstly, I_u measures how much larger ϱ_u is for a given u than the global mean self-looping tendency p. Secondly, I_u measures how much larger the latent ξ_{uu} component of $\hat{W}_{uu}$ is than its expected value of one (recall that self-loop weights are modeled as $\hat{W}_{uu} = \hat{\varrho_u} \xi_{uu} s_u$ from (3.4)).

The exact form of the variance of I_u is difficult to calculate, but we use a simulation method to approximate a p-value for I_u. We first determine estimates for a, b from the mean and variance of ϱ_u under the null model. Given p and the sample standard deviation $\hat{\sigma}_p^2$, we find estimates for a, b from (3.12). We use these estimated parameters to simulate a measurement of how extreme the $\hat{I}_u$ of a given node is, compared to that of a measurement assuming random generation from a beta$(\hat{a}, \hat{b})$ distribution. If the the actual value I_u is large, as measured by whether it is above the α-th quantile of the simulated values, then the node is deemed significant because it consistently exceeds what would be expected if ϱ_u were randomly generated from a distribution of the same

parameters. Such a simulation-based method to approximate p-values is commonly used (Ewens, 2003).

The process of finding significant *monads* may be concisely described by the following procedures. First, we simulate $\tilde{\varrho}_u$ from beta$(\hat{a}, \hat{b})$ for every node (county). We then obtain the empirical distributions of how monadic a given node is by computing the empirical distribution of $\hat{I}_u = W_{uu} - \tilde{\varrho}_u s_u$. Following this step, we consider $I_u = W_{uu} - ps_u$. If I_u is in the $1 - \alpha^{\text{th}}$ tail of the distribution $\hat{I}_u$ then the node is monadic at this instance of simulation. We repeat the procedure 10,000 times and the nodes that are monadic all of the 10,000 trials are conclusively classified as monads. In practice, we set α equal to .05 for this test. The resultant group of nodes are significantly monadic at the 5% significance level.

3.5.5 Differentiating Nodal Communities from Non-Nodal Communities

In the post-processing phase, we identify nodal communities within the communities detected by using the local clustering coefficient as defined by Opshal et al. (Opsahl and Panzarasa, 2009) for weighted networks. For an unweighted network, the local clustering coefficient of a node u is the ratio of the number of present ties over the total number of possible ties between the node's neighbors. A community will have a low clustering coefficient if there is a single hub-like node that is connected to all its peripheral nodes, but its peripheral nodes do not connect to each other. A node in a complete graph has a coefficient of 1.

For a weighted network, Opsahl et al. define the minimum clustering coefficient of a particular node u in a set C using *triplets* and *triangles* of nodes (Opsahl and Panzarasa, 2009). A triplet $\delta(u, v, w)$ is defined as a set of three nodes that share at least one edge with another node in the set. A closed triangle $\Delta(u, v, w)$ is a set of three nodes whose nodes all connect to *both* other nodes in the set. One may visually conceive of a triplet as a loosely connected set of three vertices which may have a missing edge, and a closed triangle as a clique of three vertices with three edges.

Using triplets and closed triangles, Opsahl et al. (Opsahl and Panzarasa, 2009) define the minimum clustering coefficient of node u in the set C as

$$m(u, C) = \sum_{v,y:\Delta(u,v,y)\in C} \min(W_{uv}, W_{vy}) \Big/ \sum_{v,y:\delta(u,v,y)\in C} \min(W_{uv}, W_{vy}).$$

The ratio $m(u, C)$ is the ratio of the sum of all closed triangles to triplets associated with u. If the sum of the minimum values of triangles is how compared to those of all triplets within a cluster, then that implies the presence of a dominant node that projects high edge weights across many peripheral nodes. A low clustering coefficient signifies that communities are bound together by a common node, while a high $m(u, C)$ signifies that the nodes are all connected to each other.

We determine the overall clustering coefficient of a community by

$$m_{\text{total}}(C) = \sum_{v \in C} \frac{(1 - m(v, C))s_v(C)}{\sum_{u \in C} s_u(C)}. \tag{3.20}$$

$m_{\text{total}}(C)$ is constructed to capture how tree-like the highest-weight nodes are in a given cluster C. The lower $m(u, C)$ is, the more tree-like the node is. $1 - m(u, C)$ weighted by the strengths of nodes in a given cluster assigns a value of how tree-like and strong a node is. Summing these values gives an overall measure of the monocentricity of a cluster, as the strongest nodes tend to have the smallest $m(u, C)$. The empirical distribution of $m_{\text{total}}(C)$ is bimodal (see Fig. 2 in Supporting Information) with a split around 0.4. We use this value to identify nodal communities.

3.5.6 Methods to Compare Communities with Other Delineations

We compare our results with other existing delineations (in particular OMB's Metropolitan areas) by using the Fuzzy Rand Index (FRI) (Chakraborty et al., 2017). The Fuzzy Rand Index is a metric that measures the similarity of two covers, C_1 and C_2. A cover C is the assignment of vertices of a graph into k groups, where a vertex may belong to more than one group. Counties that are not assigned to any community will belong to their own group, a group of counties that do not belong to any community. Let V represent the set of vertices and C represent a cover of V. Each element $v \in V$ is characterized by its membership vector, $C(v)$, which describes the degree of membership between a vertex and each group. A membership vector is subject to the following constraints: $C(v) = \{C_1(v), C_2(v), ..., C_k(v)\} \in [0, 1]^k$, where $C_i(v)$ is the degree of membership of v in the i^{th} community, C_i, and $\sum_{i \in 1,...,k} C_i(u) = 1$. The norm of $C(u) - C(v)$ in a cover with k communities will be defined to be $||C(u) - C(v)|| = \sum_{i \in 1,...,k} |C_i(u) - C_i(v)|/2$. For some cover C, the similarity measure between two nodes u, v is defined as $E_C(u, v) = 1 - ||C(u) - C(v)||$. The

distance measure between two covers, C_1 and C_2 of a network V is defined as:

$$d(C_1, C_2) = \frac{\sum_{u,v \in V} |E_{C_1}(u,v) - E_{C_2}(u,v)|}{k(k-1)/2}.$$

Likewise, the similarity measure between two covers is $1 - d(C_1, C_2)$.

3.6 Results

We use the term 'clusters' to refer to all sets of counties that we obtain from the detection procedures, 'communities' to refer to clusters that contain more than one county, and 'nodal communities' to refer to clusters that have strong nodal centers with little lateral commuting. We use the term 'monads' to refer to those counties that are strongly connected to themselves.

From the total 3,091 US counties, we find a total of 182 significant clusters. Of these clusters, 14 are nodal communities, 78 are non-nodal communities, and 90 are monads. Together they cover 90.3 % of the population of commuters (93% intra-county, 87% inter-county). The method simultaneously delineates both small (such as monads and dyads) and large clusters consisting of hundreds of counties. For example, Santa Barbara and San Luis Obispo counties in California are strongly connected to one another and are separate from other clusters. Of the regions identified, 99 are monadic or dyadic counties. 68 of these clusters are medium sized, comprising between 2 and 50 counties. In 20 other instances, CCME-SL identifies clusters consisting of 50 or more counties that span several states. These tend to be polycentric regions centered around multiple large cities such as Philadelphia, Washington DC, and Baltimore (see Fig. 3.2). Out of the 182 total clusters, the average size of a cluster is 24 counties, with a standard deviation of 59. The median size, however, is only 2, signifying that the majority of imputed clusters are monads. When only considering communities, the average size is 49, with a standard deviation of 78. The median size of communities is 15, suggesting that the distribution of community sizes is right-skewed (see Table **??**). In general, many of the counties belong to overlapping clusters, with some belonging to as many as six different clusters (see Fig. 3.3).

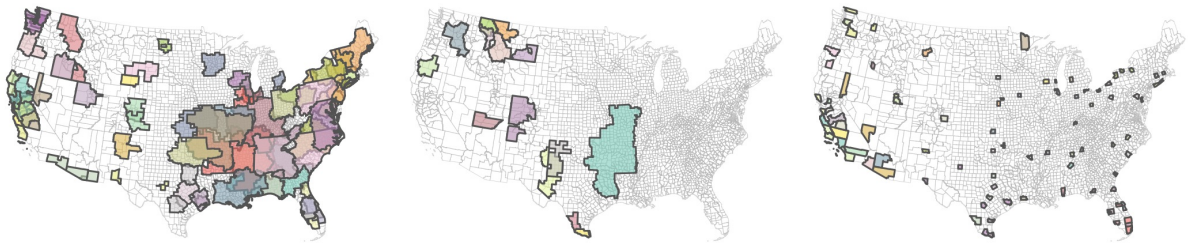


Figure 3.2: Resulting communities from the CCME-SL algorithm. Communities (non-nodal) (*left*), nodal clusters (*middle*), and monads (*right*)

3.6.1 Comparison with Other Community Detection Methods

We compare the results found by CCME-SL with several other widely used methods. We examine the results imputed by modularity maximization (Louvain) and the degree-corrected stochastic blockmodel (DC-SBM). The Louvain method naturally finds the optimal number of partitions, while the DC-SBM needs a pre-specified number of partitions. In the Louvain, DC-SBM, and *expert judgment* (OMB) methods, regions are non-overlapping, while CCME-SL is the only method that demarcates in a way that allows counties to have multiple memberships (see Fig. 3.6). A number of other techniques implicitly allow for self loops. CCME-SL was largely motivated by the need to address settings where self-loops account for a significant proportion of the weight emanating from a vertex.

The Louvain method imputes an optimal number of around 350 communities (see Fig. 3.6, bottom left) and approximately maps commuting patterns to regions roughly similar in size to (large) CBSAs. We fit DC-SBMs under two different specifications for *number of blocks*: 100, which maps approximately to the number of clusters found by CCME-SL, and 350, which are the optimally clustered sets found by modularity maximization. We visualize these clusters alongside pre-defined MSA delineations and megaregions (see Fig. 3.6).

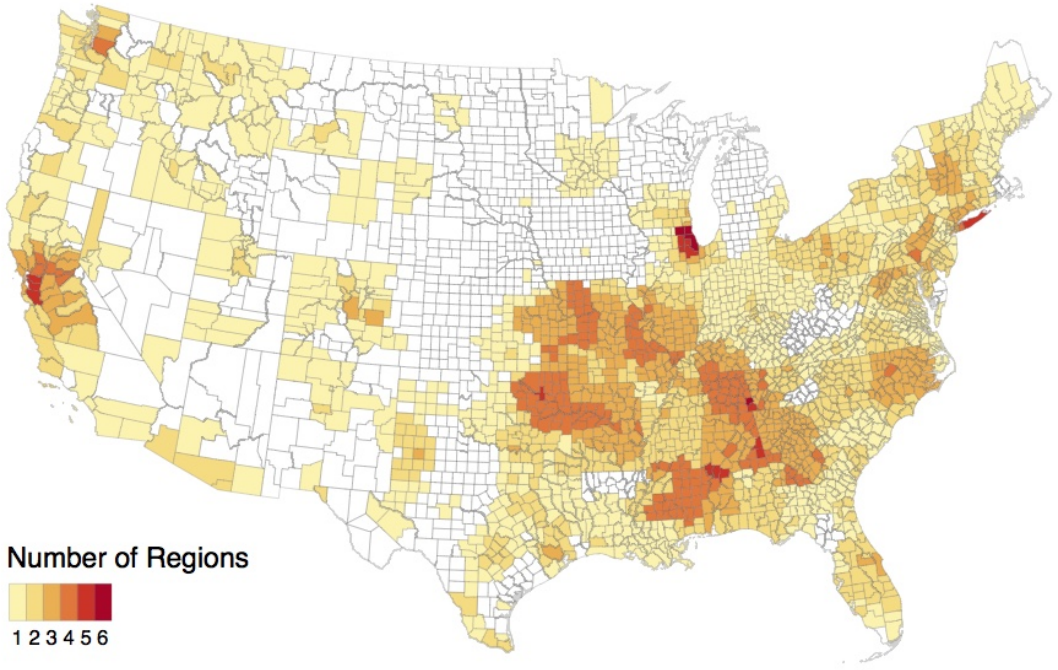


Figure 3.3: Heatmap of frequencies of each county to appear in any cluster (community, nodal cluster, or monad)

Compared to DC-SBM and modularity methods, CCME-SL is capable of capturing overlapping communities and finds much more variation in community sizes (though less so in weights). Counties that are influential for several regions, like Harris County in Texas, are strictly partitioned by DC-SBM and Louvain, but yield components in both 'coastal' and 'inland' counties in communities found by CCME-SL. Communities imputed by DC-SBM are highly dependent on the pre-specified number of blocks chosen. Los Angeles County is a single block under DC-SBM when 100 counties are chosen, but is subsumed by a much larger block when 350 blocks are chosen.

DC-SBM was implemented by means of regularized spherical spectral clustering (Qin and Rohe, 2013), which has been shown to be consistent with DC-SBM in the package *randnet*. The modularity algorithm was implemented using the package *igraph*.

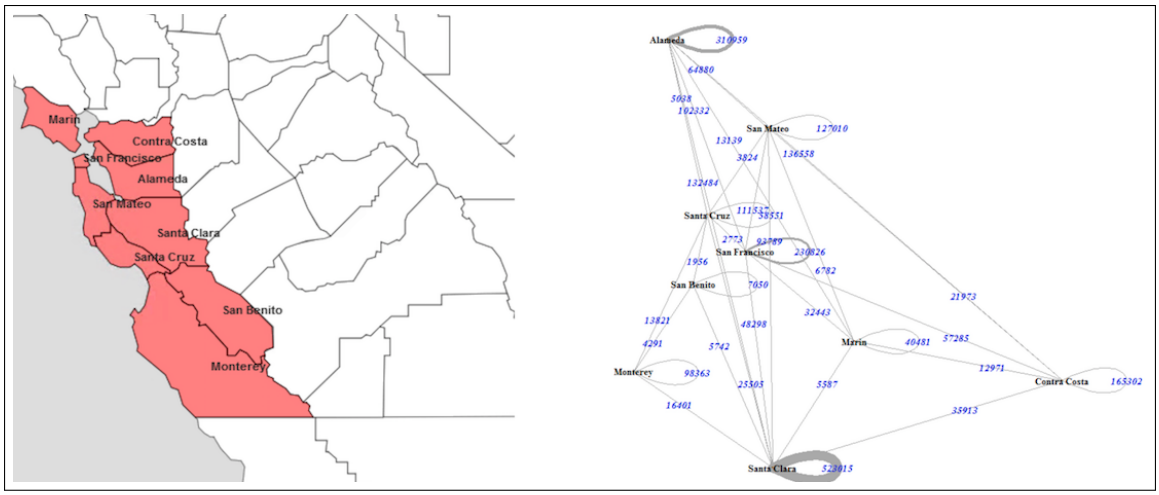


Figure 3.4: Example of a tightly connected community in the Bay Area in Northern California

3.7 Discussion

We summarize our findings in this section and offer interpretations in relation to the economic geography of the US. We highlight how and why our findings are different from typical delineations and reconcile these findings with the introduced method and its novel incorporation of self-loops in a null model within a weighted network.

CCME-SL produces clusters that vary greatly in size. The most populous MSAs house similar counties as their corresponding communities (see Fig. 3.5). However, communities can also be as large as megaregions, although they tend to capture counties in different ways. Megaregions capture cities that are in close proximity and which have large overall commuting volumes,but communities capture sets of counties that are closely linked by commuting, even when there are no cities and the gross commuting volume is not large. In the South and Central parts of the US, counties tend to be small and rural yet tightly interconnected. Such aggregations have not been depicted in existing official regional delineations and may be a novel contribution of the method in this study.

3.7.1 Methodological Contributions to Region Demarcation

Though detection of monads may be antithetical to typical notions of *community* in network theory, they are very important in this particular application. This research highlights the role of self loops in commuting networks and is broadly applicable to human mobility networks with

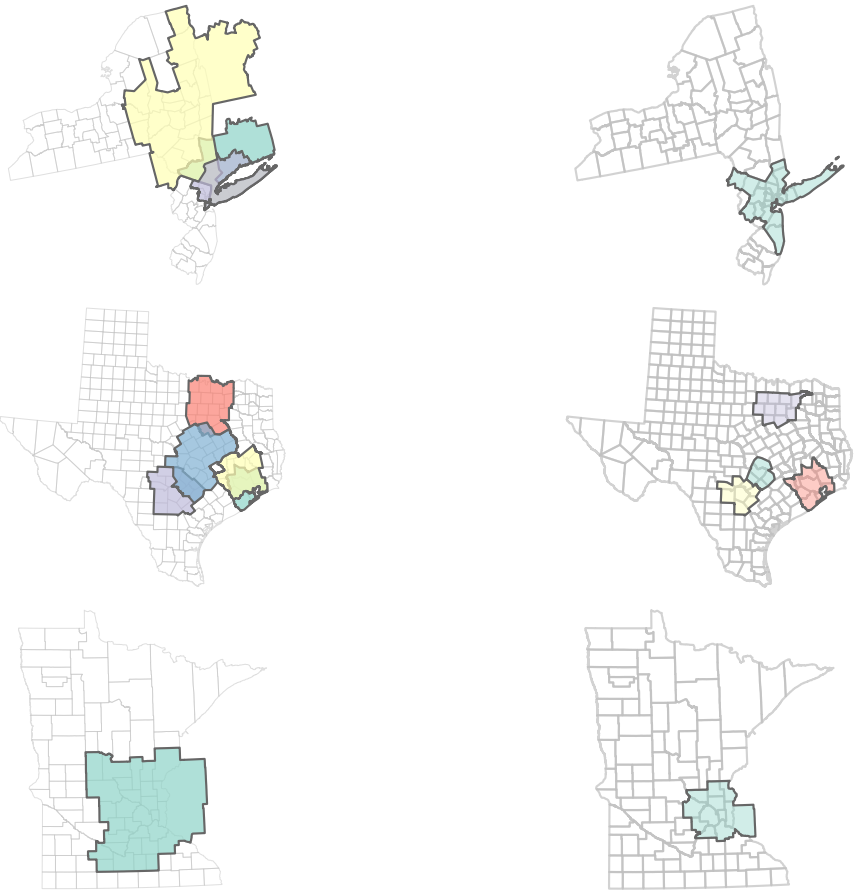
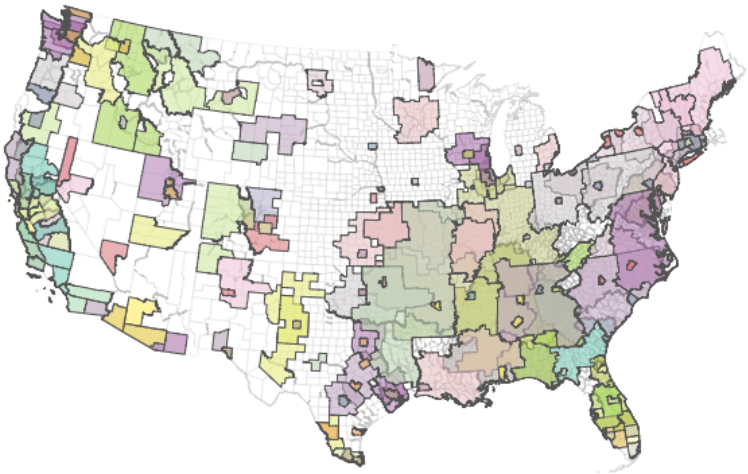


Figure 3.5: Comparison of MSAs of New York City Region, major Texas cities, and Minneapolis (left) with their associated communities (right) in fairly populous regions

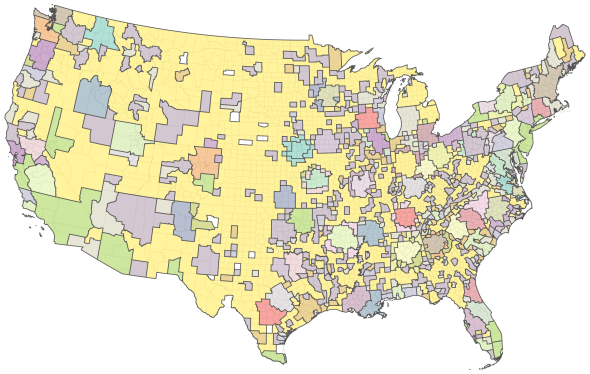
spatial constraints. As described in (Clauset et al., 2009), most network data arising from nature adhere to power laws, and commuting flows are no exception. This study shows that power laws in spatial settings are intricately linked to self-referential behavior. Many studies have described human populations adhering to heavy-tailed distributions such as the Zipf Law (Newman, 2005; Barabási and Albert, 1999), but this study is among the first to indicate how regional delineations can account for such phenomena.

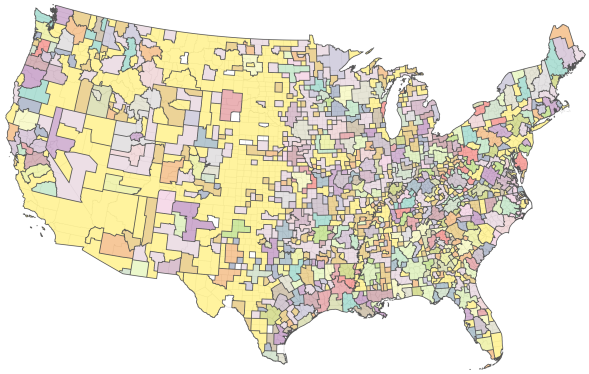
Strongly self-commuting counties that are identified as monads are also classified as clusters. Monads are found using tests of similar hypotheses evaluating how unlikely *the size of the total commuting activity in a given geographical unit is* compared to a scenario where commuting activity was generated at random subject to constraints arising from the weighted configuration model. The null hypothesis for monadicity parallels the null hypothesis for community connectivity in evaluating whether a test node is significantly more strongly connected to other nodes, or itself, than expected.

CCME-SL

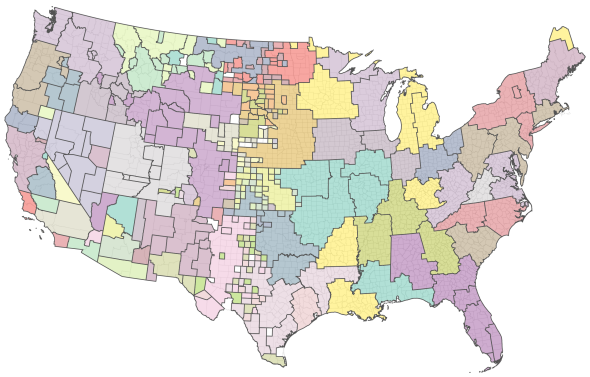


Degree-Corrected Stochastic Blockmodel (100 Blocks (L), 350 Blocks (R))





Modularity Algorithm (Louvain)



MSA and Megaregions

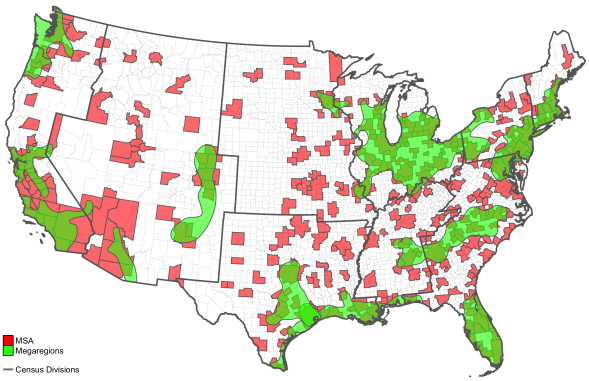


Figure 3.6: Comparison of clusters from of CCME-SL (top) with DC-SBM (middle row, 100 blocks (left), 350 blocks (right)) , Modularity (bottom left) , and MSAs and megaregions (bottom right)

CCME-SL is mostly unsupervised. There are only two tuning parameters controlling the algorithm, α and τ. α tunes the significance of the nodes' connectivity in relation to its associated community, and τ tunes the threshold of Jaccard distance to filter overlapping clusters. Different values of α and τ are used for both communities and monads, the results of which are shown in the Supporting Information. We set the $\alpha = 0.01$ and $\tau = 0.80$ for communities and $\alpha = 0.05$ for monads because monads cover nearly all the major US metropolitan areas. Though they don't differ much from those presented in Section 3.6, communities induced by different thresholds promote discovery of more tightly linked or more significantly *monadic* clusters.

Nelson and Rae use a multi-step routine to tune parameters for modularity-based methods, filtering out unwanted 'outlier' nodes and validating visual heuristics (Nelson and Rae, 2016). Compared to this approach, our method is more parsimonious and statistically interpretable.

3.7.2 Comparisons with Other Community Detection Methods

In Section 3.6.1, we compared the CCME-SL algorithm with several other standard community detection techniques, including the existing OMB demarcations. The primary advantage of CCME-SL is that it accounts for overlapping memberships for each node. Another advantage of CCME-SL is that the regions demarcated are more defensible because the model accounts for self-loops, which are important in the commuting network. Finally, the simplicity of CCME-SL is a practical advantage compared to other models, especially DC-SBM, which requires that the optimal clustering parameters be determined through cross-validation.

Though other techniques *implicitly account for* self-loops, the self-loops oftentimes cause distortions that create problems in identifying regions. Fig. 3.6 shows that Louvain and DC-SBM yield smaller communities that are all roughly balanced in number of nodes per community. Sizes of regions are highly contingent on their populations and commuting volumes. CCME-SL, on the other hand, yields clusters that are highly variable in size, commensurate with the highly variable populations. The sizes and characteristics of the clusters (monads and communities) imputed by CCME-SL thus appear to be *less constrained* than the other three methods shown in Fig. 3.6. We posit that the differences between clusters from CCME-SL and other approaches are at least partly due to the extremely strong self-looping weights in high-strength nodes (populous counties).

Many algorithms return different results under different initialization scenarios when the likelihoods of partition functions are multimodal and thus give rise to a number of near-optimal partitions (Good et al., 2010; Peel et al., 2017). Though this is a common problem in modularity-based approaches with random seedings, CCME-SL is not strongly affected by this issue for two different reasons. First, the method of initialization using the heuristics of starting at nodes with self-loops larger than 20,000 described in Section 3.5.1 produces the same results upon every run of the algorithm. Second, even if initialization was randomized and subject to different initialization criteria (while still retaining most of the population-center counties), the results do not look very different (see Fig. 4 in Supporting Information). In other applications of CCME-SL, wherein the data do not have interpretable initial seeds, more runs would be required and analysis of *sets of partitions* would be necessary (Peixoto, 2017).

The number of communities is very important in community detection and its determination is a difficult problem in the field. For example, in DC-SBM the number can be determined by model selection or by cross validation. Under CCME-SL, the number of detected communities is determined by just one sample. This would be just one of the near-optimal states of the assumed model, as there would be other optimal states with different numbers of communities. Although the proposed method finds generally similar communities under a range of parameters and initializations (Figs. 3,4 in Supporting Information), these validations do not allow for the discovery of some *exact* optimal objective. As such, we reiterate the point that CCME-SL should be viewed as an exploratory method that could give rise to more rigorous modes for proposing novel OMB-designated regions based on the structure of a commuting network.

Traditional delineations of geographic regions have relied on agglomerations of smaller geographies, historical and political boundaries, separating edges and central foci. The boundary characterization is important not only for scientific purposes of tracking and tracing the historical evolution of urban systems, but also for administrative purposes of allocating infrastructure investments and formulating economic development strategies. Boundaries of metropolitan areas in the United States are artifacts of these delineation definitions, yet are central to tracking demographic and economic changes, funding allocations, determination of fair market rents, housing subsidies that depend on area median income and a host of other federal and state programs, even when the agencies caution their use for non-statistical purposes. These delineations are central but invisible

to the lives of many. In this chapter, we provide a robust method of accounting for the membership of a single place in multiple regions.

The main methodological contribution of this chapter is its introduction of a community extraction method for a network with strongly self-looping characteristics. The application of the method on US commuting data suggests a way of conceiving delineations of economic geography that differs from existing approaches. CCME-SL accounts for intra-county commuting patterns and produces drastically different results when compared to other community detection methods as well as CBSA-based approaches. Furthermore, allowing regions to overlap allows us to create institutional structures and policies that are tailored not only to singular geographical entities, but also to multitudinous identities interacting across space and place.

CHAPTER 4

Intertemporal Community Detection in Human Mobility Networks

Much research has been done in recent years in the analysis of real world networks. One particular area of active interest is in *intertemporal community detection*[1]. A majority of the research on community detection in networks has dealt with static networks (Girvan and Newman, 2002). However, many real-world networks exhibit dynamic properties, such as human mobility networks in urban systems. These networks include commuting patterns over time (Patuelli et al., 2010), location based social networks (Assem et al., 2016), taxicab travel patterns (Liu et al., 2015) and cell phone call records (Reades et al., 2009). Understanding the structures of these networks reveals underlying trends in human mobility and provides important information for the management of urban infrastructure.

There are many human mobility patterns that can be represented as networks with high temporal resolution because of the presence of origin and destination locations and time stamps associated with the trips. For example, bikeshare systems are rich and remarkably comprehensive in tracking mobility patterns within a city. By 2019, over 2000 cities have created bikeshare systems around the world. In 2018, according to the National Association of City Transportation Officials, 36.5 million trips were completed in over 100 cities in the United States using these systems. Many of these systems have stations where users can rent the bikes and deposit them at another station at the end of the trip. These stations allow the system operator to track the precise origins and destinations of individual trips by time-of-day and day-of- week. Travel by automobile can be modeled as networks: in particular, taxicabs in cities are regulated and therefore location and time data of these cab pickup and dropoff locations are often reported to the regulators. The increased usage of often less-regulated ridesharing services (Uber, Lyft etc.) have reduced taxi trips in the last few years. Much research has been done on network analyses (Austwick et al., 2013; Cazabet

[1]This chapter is adapted from a manuscript written in 2020 (He et al., 2020a), joint work with Professors Shankar Bhamidi and Nikhil Kaza

et al., 2017b; Zhan et al., 2016; Tong et al., 2017), but most do not fully take into account the dependencies induced by the network structures and temporal trends. Many of these studies have also focused mostly on demand estimation (Zhou, 2015; Faghih-Imani and Eluru, 2015).

In this chapter, we develop a method to identify clusters of significantly connected nodes in a time-series of weighted networks. Identification of such clusters allows us to understand the nature of geographical, economic and cultural relationships, when networks are representative of urban systems. Identifying trajectories of connectivity in clusters across time may reveal structural changes within the mobility patterns in these systems. We develop an intertemporal community detection method to analyze the structure of long-term trends in time-series of networks to understand *global* and *local* trends. In particular, we attempt to determine whether such trends are uniformly distributed across the networks, or whether certain communities exhibit countervailing trends in interconnectivity when compared with others. We aim to identify and partition the nodes that are part of communities which exhibit locally specific trends.

The objective of the community detection method in this study is to find groups of nodes that are consistently connected across time and exhibit increasing, decreasing, or stable trends in connectivity. Our methodological framework rests on the assumption of a baseline null model that preserves the functionals of the observed network, then extracting subsets of vertices that exhibit significant deviations in connectivity contrasted with the null model. We use a weighted configuration model as posited in (Palowitch et al., 2018; He et al., 2020b). We extend the framework and introduce additional steps to find consistent patterns of connectivities among clusters across time.

Analysis of time-varying weighted graphs allows us to gain more insight into the nature of the city as a complex accumulation of micro-level spatial activity patterns. While this method of intertemporal community detection is developed for data structured like mobility systems, it can be adapted for any type of time-series network data with registered nodes (such as inter county commuting patterns, internet traffic, etc.).

4.1 Layout and Contributions

The primary contribution of this chapter is in proposing a novel community detection method in human mobility networks. Though much work has been done on community detection in single

(static) networks, in the realm of time-varying networks there are still many open avenues in the realm of clustering. When the time-series are structured in mid-to-high frequencies, many existing methods of multilayer community detection run into identifiability issues, particularly those using model-based approaches (Matias and Miele, 2017). By presuming simplified trajectories of connectivity to *increasing*, *decreasing*, and *neutral* and by using iterative testing techniques , the method is a fast, reliable, and parsimonious way to discover overlapping clusters in high complex interrelational datasets.

In this chapter, we first describe the procurement and preprocessing of bikeshare and taxicab data in section 4.2. We then describe the methodology in the following section 4.3, starting from the description of the configuration null model at a single time-slice in subsection 4.3.1, then progressing to describe the time-varying FDR correction of the significance between bordering nodes to sets from sections 4.3.3 to 4.3.5. We then describe the testing for significance of the trends of connectivities in section 4.3.6 and finally describe the initializing and overlap-filteirng steps. In the following section 4.5 , we describe the resultant clusters and discuss their potential interpretations in section 4.6. In section 4.4 we detail additional methodology for estimating foregone trips within bikeshare networks due to load imbalance.

4.2 Data and Network Construction

We apply intertemporal community detection to data from two bikeshare systems and a taxicab trips. Bikeshare trip data for Divvy (Chicago) and Citibike (New York) are publicly available on their respective websites (Divvy, 2019; Citibike, 2019). The two bikeshare systems provide contrasting cases. Divvy ridership increased steadily between 2014-2016 from 2.7 to 3.6 million, but overall ridership declined slightly from 3.8 million trips in 2017 to ∼3.6 million in 2018 (Greenfield, 2018). The Citibike system, on the other hand, has consistently increased in usage from 14 million in 2016 to 16 million in 2017 and 18 million in 2018 (Citibike, 2019).

The publicly available datasets include trip start and stop times for each trip between stations. In our analyses, we focus on the time period between July 2016 and June 2018. We omit all stations that were newly introduced or removed within this period. There remained 547 nodes (7.4 million trips) in Chicago and 583 nodes (8.4 million trips) in New York in the dataset used for this study.

One common problem in bikeshare systems is the issue of *supply-demand mismatch* in ridership. A station in a high-activity area of a large city is often empty or full at certain times of the day (Gast et al., 2015; Xie and Wang, 2018; Pendem, 2019; Freund et al., 2018; Faghih-Imani, 2014; Zhou, 2015). A full or empty station prevents an otherwise possible trip. Load rebalancing is a well-studied problem for bikeshare systems in order to solve the inefficiencies associated with queuing between bikes in stations with finite numbers of slots for bikes at each station. Real-time data on station status rebalancing exist for New York and Chicago (Divvy, 2019). However, historical station inventory data is only available for New York City (Open BUS, 2019) and not Chicago. Thus, for the New York bikeshare system, we find communities with and without demand adjustment (see section 4.4 for details on the method).

The taxicab data for New York is from the Taxicab and Limousine Commission (NYC Taxi and Limousine Commission, 2020). We use data from January 2017 to 2019 because trips from the ridehailing apps (such as Uber, Jio) are only included since 2017 in the data. There are 263 pick-up/dropoff zones, which cover all the five boroughs of New York, and the dataset includes over 453 million trips between these zones.

From these datasets, we construct the observed time-series of networks as $\{G_t\}_{1\leq t\leq T}$. In all these datasets, we aggregate the trips between a pair of nodes for each week. The weekly aggregation smooths the diurnal variations and keeps the time-series long enough for time-domain analysis. Thus, each time t corresponds to a week, where T is the total number of time periods. The indicator $A_{uv,t}$ represents the presence of any trips at time t between u and v. We use the number of trips between two nodes at week t as the edge weight $W_{uv,t}$. In network G_t the degree of node u is defined as $\deg_{u,t} = \sum_{v:v\neq u} A_{uv,t}$ and strengths are defined as $S_{u,t} = \sum_{v:v\neq u} W_{uv,t}$ at each time-unit t across total time T (He et al., 2020b). We de the index set $[n] = \{1, 2, ..., n\}$ as the set of all nodes u, which represent stations in bikeshare systems and pick-up/dropoff zones for taxicab networks.

4.3 Detecting Intertemporal Communities

In this section, we describe a method to extract statistically significant communities across time (He et al., 2020b; Palowitch et al., 2018) based on iterative testing of node-set connectivities. We

use a similar approach but account for and classify the types of time dependency. We posit that trends across time are generally *increasing, decreasing* or *stable* and account for these types of time dependence. To this end, we adjust connectivities to time-decay and find trends using equivalence testing (Schuirmann, 1987; Dixon and Pechmann, 2008).

4.3.1 Intertemporal Configuration Null Model

We use a similar epistemological heuristic as in (Palowitch et al., 2018; He et al., 2020b) in positing a baseline model that preserves the characteristics of the observed network,then extracting subsets of vertices that exhibit significant deviations in connectivity contrasted with the null model. The framework of the method posits a baseline model is extracted from a time-series of registered networks, which are then iteratively subjected to hypothesis tests for trends and local deviance. We detect significant communities across the time-series of networks if the trend **and** variation components are significantly different from those of the baseline model. These communities signal subsections of the network that are either strongly interconnected at either the beginning or end of the time-period, or consistently connected throughout the entire time period.

The intertemporal null model for a given node set B (as in (Palowitch et al., 2018; He et al., 2020b)) is defined to determine if it is significantly interconnected across *all* time-points according to the hypothesized trend. We search for communities that are

- *decreasing* if its nodes are significantly connected at time $t = 1$, but not necessarily significantly connected as t becomes larger (later time period), such that nodes that are significantly connected in the beginning, but not at the end, are identified.
- *increasing* if its nodes are significantly connected at later times (when t approaches T), but not necessarily significantly connected when t is early.
- *stable* (or neutral) if its nodes are significantly connected across all time points.

If vertices are not connected at all time points, then they are not clustered and left in the "background". Like in (Palowitch et al., 2018), vertices can belong to more than one cluster. Within B, a time-series of relative connectivity may be decomposed into *trend* and *variation* components. Trend denotes the presence of a constant time-trend in the relative connectivity amongst nodes in

set B. Variation denotes the aspects of the node-set connectivity that do not vary systematically across time.

4.3.2 Null Model for Node-Set Connectivity

The estimate for each edge weight $W_{uv,t}$ at time t is a simple extension of the model used in (Palowitch et al., 2018), which is the null model for a single graph.

$$\widehat{W}_{uv,t} = \begin{cases} \xi_{uv,t}\left(\frac{s_{u,t}s_{v,t}}{s_{T,t}}\right)\Big/\left(\frac{d_{u,t}d_{v,t}}{d_{T,t}}\right) & \text{if } u \neq v \\ 0 & \text{if } u = v \end{cases} \tag{4.1}$$

Each $\widehat{W}_{uv,t}$ is a weighted edge on a random time-varying graph $\mathcal{G}_t$, where each u has fixed degrees $d_{u,t}$ and strengths $s_{u,t}$. Each graph G_t at time t has total degrees $d_{T,t} = \sum_v d_{v,t}$ and total strengths $s_{T,t} = \sum_v s_{v,t}$. Random variables $\xi_{uv,t}$ with mean 1 and variance κ_t are constructed so as to satisfy the weighted configuration model used in the work of Palowitch et al. (He et al., 2020b; Palowitch et al., 2018). The analogous node-set connectivity $S(u, B, G_t)$ (as in (Palowitch et al., 2018)), is

$$S(u, B, G_t) = \sum_{v \neq u, v \in B} W_{uv,t}. \tag{4.2}$$

This value measures how each node u at connects with the set of nodes B at time point t. Each score $S(u, B, G_t)$ is fixed across a given node u and set B, but is different for every time-step t. A central limit theorem is used in (Palowitch et al., 2018; He et al., 2020b) to approximate $S(u, B, G_t)$ as a normal distribution with means and variances

$$\mathbb{E}[S(u, B, \mathcal{G}_t)] = \sum_{v \in B} \frac{s_{u,t}s_{v,t}}{s_{T,t}}; \tag{4.3}$$

$$\text{Var}(S(u, B, \mathcal{G}_t)) = \sum_{v \in B} \frac{\left(\frac{s_{u,t}s_{v,t}}{s_{T,t}}\right)^2}{\frac{d_{u,t}d_{v,t}}{d_{T,t}}} \left(\kappa_t - \frac{d_{u,t}d_{v,t}}{d_{T,t}} + 1\right). \tag{4.4}$$

A p-value is derived from the above statistics in order to gauge the probability of the node-set connecitivity as significantly deviant from what it would be under the null model. P-values are

derived from the normalized test statistic $Z_t(v, B)$, defined below:

$$Z_t(v, B) = \frac{S(v, B, G_t) - \mathbb{E}[S(v, B, \mathcal{G}_t)]}{\sqrt{\text{Var}(S(v, B, \mathcal{G}_t))}}, \quad t = 1, ..., T. \tag{4.5}$$

$Z_t(v, B)$ is posited to follow a $N(0, 1)$ distribution and values that significantly exceed the distribution under the null model are identified to be significantly connected. sets of nodes that are all significantly connected, statistic $Z_t(v, B)$ is computed for every node $v \in [n]$ at time t. Significantly connected nodes are extracted with respect to a given set B in order to identify sets of nodes wherein every member is significantly connected to each other in the set. Significance in this case is determined by p-values which are calculated as follows:

$$p(u, B, G_t) = \mathbb{P}(S(u, B, G_t) > S(u, B, \mathcal{G}_t)). \tag{4.6}$$

Nodes with respect to B are augmented by false-discovery-rate corrrections and selected based on a pre-specified signifiance threshold α, which conventionally is equal to or below 0.05 (95 % significance). In practice, these p-values are iteratively computed several times until the membership of the set converges. Detailed derivations of these values can be found in the text of (Palowitch et al., 2018).

4.3.3 Identifying Nodes that are Significantly Bordering Across Time

We use iterative testing to identify nodes that are significantly connected to their neighbors through time. Methods developed in previous literature (Palowitch et al., 2018; He et al., 2020b) have applied this method to a fixed graph G. We use the same method of deriving significance of the probability that v is significantly connected to u in set B as in those methods.

Our proposed method relies on an iterative procedure starting at iteration step $k = 1$, then repeated until the results do not change. The objective is to find sets B such that for each $v \in B$, v is significantly connected to u across *all* time points $1, .., T$. At a given step $k > 1$, for fixed time t, for a set of nodes $B_{k,t}$ and a bordering node u, the score of node-set connectivity is determined by (4.2):

$$S(u, B_{k,t}, G_t) = \sum_{v \neq u, v \in B_{k,t}} W_{uv,t}. \tag{4.7}$$

After the normalizing calculation (4.5) is performed, a p-value for each $v \in B_{k,t}$ is then determined as in (4.6)

For each time point t, the p-value $p(u, B_{k,t}, G_t)$ is then corrected for false-discovery rate correction as in (Wilson et al., 2014). The non-significant nodes are rejected and the set of significant nodes is retained. Additional steps to find significant nodes are described in the following sections 4.3.4 - 4.3.5 to account for time-decay in significant bordering nodes and describe the testing of trends in 4.3.6.

4.3.4 Time-Decay Adjusted False Discovery Rate Correction

To identify significantly interconnected nodes for a given time t, an augmented version of the Benjamini-Hochberg (Benjamini and Hochberg, 1995) procedure is used. The BH procedure is used in (Palowitch et al., 2018; He et al., 2020b) , but the difference in this approach is that the FDR-adjusted p-value p_u^* is multiplied by decay term a_t, contingent on if the communities are hypothesized to be increasing, decreasing, or stable in connectivity over time. For a fixed time t, iteration step k, and set $B_{k,t}$, we find all the nodes that are significantly connected to $B_{k,t}$ across all time $t = 1, ..., T$ after calculating the p-value as in (4.6). The output set at iteration K and time t is written as $M_k(B_k)$, described in more detail in later sections in equation (4.10).

We define a_t is an exponential decay term to adjust for the shifting time-window of significance. It is defined as:

$$a_t := \begin{cases} \left(1 - \exp\left(-\frac{t-1}{T}\right)\right) a_0^+ & \text{if trend is increasing} \\ \left(\exp\left(-\frac{t-1}{T}\right) - a_0^-\right) \Big/ (1 - a_0^-) & \text{if trend is decreasing} \\ 1 & \text{if trend is neutral.} \end{cases} \tag{4.8}$$

The terms a_0^+ and a_0^- are defined such that a_t is 0 at time 1 and 1 at time T if the trend is increasing, and 1 at time 1 and 0 at time T if the trend is decreasing:

$$a_0^+ := 1 - \exp\left(-\frac{T-1}{T}\right); \qquad a_0^- := \exp\left(-\frac{T-1}{T}\right).$$

If the trend is posited to be decreasing, then the algorithm allows more permissive selection of 'significantly' bordering nodes when time t is early, but is more penalizing when t approaches T. When t is 1, then a_t is equal to zero. In this case, all p_u^* are zero and will automatically be counted as significant if u borders $B_{k,t}$. When t is T, then a_t is 1, so the FDR correction is identical to BH. The threshold for the maximum allowable p-value increases as t decreases so that negligible connections (when t is early) that become stronger (when t is late) are deemed significant.

Conversely, when the trend is posited to be decreasing, the same kind of adjustment is made in reverse because the multiplier is subtracted by one. Because the multiplier to the adjusted p-value is always less than 1, the procedure is always less conservative than the Benjamini-Hochberg method and allows nodes that otherwise would not be significant at a given time-period be deemed as"significant" based on their potential to be significant given their trajectory. If the trend is posited to be neutral, then we use the ordinary BH rejection procedure.

4.3.5 Bonferroni Interval for Bordering Frequencies

The previous section 4.3.4 details significance testing for the collections of nodes $B_{k,t}$ at each time period. To determine whether the collections of nodes are significantly connected to u at *all* times, we apply a second testing step using Bonferroni Correction. This correction is applied to the frequencies of nodes whose p-values have been deemed significant by the BH correction (in the previous section). The product of Bonferroni confidence intervals is used to define the significance of the neighboring frequency at iteration step k, for each $B_{k,t}$ across all time $t = 1, ..., T$. For a set $B_{k,t}$, we define $m_t(B_{k,t})$ as the set of nodes that are found to be 'significantly bordering' described by Section 4.3.4:

$$m_t(B_{k,t}) = \#\{u\text{: } u \text{ is significantly bordering } B_{k,t} \text{ at time } t\}.$$

A large value of $m_t(B_{k,t})$ for all t signifies a large collection of nodes that significantly border $B_{k,t}$ and results in a false discovery interval that is close to T, and hence v must border $B_{k,t}$ for nearly all time T for it to be significant. Conversely, if $m_t(B_{k,t})$ is small, then the required frequency for v to be significant is not as high. During this step, we assume away dependency between $B_{k,t}$.

We define $FDI_{\alpha,k}$ to be the threshold for false discovery interval of all significantly adjacent nodes to node $B_{k,t}$

$$FDI_{\alpha,k} = \prod_{t=1,...,T} \left(1 - \frac{\alpha}{m_t(B_{k,t})}\right) \cdot T$$

where $1 - \alpha/m_t(B_{k,t})$ is the Bonferroni confidence level at each time point . The product of these intervals cross all time multiplied by the total time T gives the threshold of significantly bordering nodes across all time.

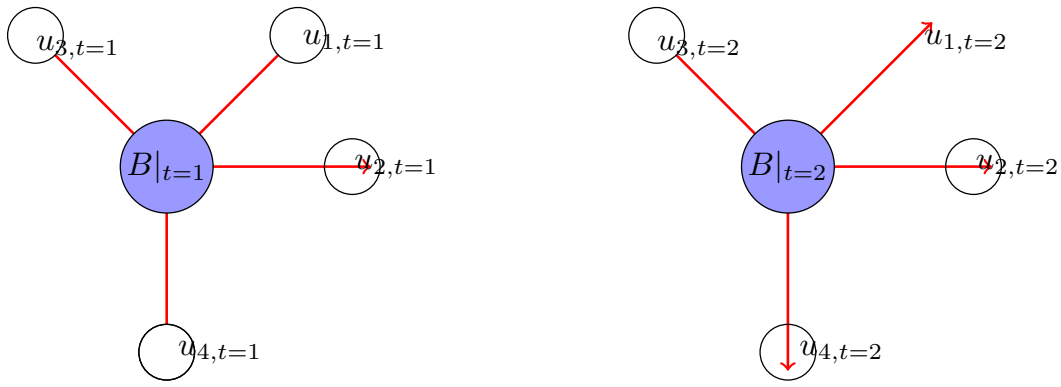


Figure 4.1: Example of set B at times $t = 1, 2$. u_1 is significantly connected when $t = 1$, but not when $t = 2$. So for arbitrary iteration step k, let $B_k = B$, then $m_t(B_{k,t})$ is $m_1(B_{1,k}) = B_k \bigcup \{u_1, u_2, u_3, u_4\}$ at $t = 1$, but $m_2(B_{2,k}) = B_k \bigcup \{u_2, u_3, u_4\}$ at $t = 2$.

Now we define the B_k^0 as the combined list of all the nodes in any $B_{k,t}$: $B_k^0 = \bigcup_{t=1,...,T} B_{k,t}$. For each $v \in B_k^0$, we define the bordering frequency $N_v(B_k)$ as the counts of v which are significantly bordering $B_{k,t}$ across all time t. A significant $N_v(B_k)$ suggests that v is more frequently bordered across time than other nodes. Each v significantly borders all $B_{k,t}$ if

$$FDI_{\alpha,k} < N_v(B_{k,t}) \tag{4.9}$$

that is, if $B_{k,t}$ borders v enough times across t for it to be significant overall in the time-period $1, ..., T$ (Dunn, 1959). Finally, we take the union of all nodes v that satisfy the "significantly neighboring" criteria (4.9) and denote the set $M_k(B_k)$

$$M_k(B_k) = \bigcup_{v \in B_k^0} \{v : FDI_{\alpha,k} < N_v(B_k)\}. \tag{4.10}$$

The resulting set $M_k(B_k)$ represents the nodes that are significantly connected across time, given the appropriate time-window adjustments. We then check if the trends are actually as hypothesized.

4.3.6 Significance Testing for Trends

We define the sum of $Z_t(v, B)$ in (4.5) as $\mathbf{Z}(B)$ to gauge the significance of the time-trend of a cluster.

$$\begin{aligned}\mathbf{Z}(B) &= \left\{\sum_{v\in B} Z_t(v, B)\right\}_{1\leq t\leq T} \\ &:= \mathbf{V}(B) + \sum_{v\in B} \beta_{v,B}\mathbf{t}, \end{aligned} \tag{4.11}$$

Moreover, for a given community B that is significantly connected across time $t = 1, ..., T$, we write the vector of node-set connectivity $\mathbf{Z}(B)$ as the sum of *trend* and *variation* components, where $\beta_{u,B}\mathbf{t}$ represents the trend component which is linearly dependent on time and $\mathbf{V}(B)$ represents the variation component that is stationary across time.

The previous sections describe discovery of node-sets that are significantly connected across time, this section details testing for their trends. If the trends are posited to be positive or negative, then one-sided t-tests are used, respectively with null hypotheses $H_{0,+} : \beta_{v,B} \leq 0$ and $H_{0,-} : \beta_{v,B} \geq 0$. If the the trend is posited to be negligible (stable), then the two sided test:

$$H_0 : \beta_{v,B} \neq 0; \qquad H_1 : \beta_{v,B} = 0$$

is used. The hypothesis is flipped (compared to the positive or negative tests) in order to test if the trend is equal to zero. We invoke equivalence testing methods ((Dixon and Pechmann, 2008)) to determine significance in relation to a pre-selected symmetric interval $[-U, U]$ about zero.

Given an set B_k at iteration k, we first find all nodes v^* that are *significantly bordering across time* as described in Section 4.3.3 and label these nodes as $M_k(B_k)$ as in (4.10). We then assess the significance of the trends of each of the nodes $v \in M_k(B_k)$ in relation to set B_k. Calculation of

trend employs test statistic for node-set connectivity $S(u, B_k, G_t)$:

$$\mathbf{Z}(v, B_k) = \left\{ \frac{S(v, B_k, G_t) - \mathbb{E}[S(v, B_k, \mathcal{G}_t)]}{\mathrm{Var}(S(v, B_k, \mathcal{G}_t))} \right\}_{1 \leq t \leq T}. \tag{4.12}$$

Using B_k and $M_k(B_k)$, we then find the time trend β_{v,B_k} for each $v \in M_k(B_k)$. We assume that intertemporal communities have trends that are *increasing*, *decreasing*, or *neutral*. We use the equivalence testing method to assess trend significance (Schuirmann, 1987; Dixon and Pechmann, 2008) . Even if a trend is significant, its impact may be negligible and should be assumed to be "zero". A bounding energy barrier $U > 0$ is chosen to control the size of the desired time-trends. A positive U is chosen as a lower bound for a positive trend, $-U$ is used as a upper bound for a negative trend. A symmetric bounding interval of $[-U, U]$ about zero is used for a neutral trend.

Hypothesis tests are conducted for the time trend for set B_k (at iteration k) and node v. Significances of trend $\beta_{v,B}$ (assuming fixed $B := B_k$ at iteration k) are calculated using the difference of the estimates with the upper bounds U (if positive) and lower bound $-U$ (if negative). T-tests for these differences $\beta_{v,B} - U$ or $\beta_{v,B} + U$ are then performed to assess significance while excluding very small trends. To determine whether a node-set has a significantly negligible (neutral) trend, we utilize the approach outlined by Dixon et al. (Dixon and Pechmann, 2008) and use two one-sided tests to determine if $\beta_{v,B}$ is significantly outside the interval $[-U, U]$. Details on the test statistics can be found in the following section 4.3.7.

4.3.7 Testing for Increasing and Decreasing Trends among Node-Sets

For the time trend expressed w.r.t. t given a set B, node v, we test for hypotheses for trend about a symmetric interval $[-U, U]$ close to zero. These hypotheses test for a null hypothesis of zero in equivalence testing. The null hypotheses are written as follows:

$$H_{0,+} : \beta^+_{v,B} \leq U \quad H_{1,+} : \beta^+_{v,B} > U, \tag{4.13}$$

$$H_{0,-} : \beta^-_{v,B} \geq -U \quad H_{1,-} : \beta^-_{v,B} < -U. \tag{4.14}$$

We calculate the significance of $\beta_{v,B}$ using the difference of the estimates as well as the (pre-specified) upper and lower bounds of the trend. In order to filter out the trends that are negligible,

we perform a t-test for the regression statistic subtracted by the upper or lower bound U, divided by the standard error of the estimate, $s(\beta_{v,B})$. Defining such a bound allows us to exclude the very small but still significant trends and only find clusters that are increasing or decreasing with considerable magnitude.

$$t_{\text{upper}}(v,B) = \frac{\hat{\beta}^{+}_{vB} - U}{s(\beta^{+}_{vB})}, \qquad t_{\text{lower}}(v,B) = -\frac{\hat{\beta}^{-}_{vB} - (-U)}{s(\beta^{-}_{vB})}$$

The corresponding p-values of t_{upper} and t_{lower}, respectively, with significance $\alpha/2$ (for one-sided tests) and with degrees of freedom $n-2$, represent the trend of connectivity of node v in relation to set B. Typical of ordinary least squares, the degrees of freedom are discounted by the slope and intercept terms.

P-values of the similarity of neutral trends to U are obtained by taking the maximum of the p-values associated with the t-statistics $t_{\text{neutral},a}$ and $t_{\text{neutral},b}$, respectively, with significance α and degrees of freedom $n-2$.

To determine the t-statistic of a negligible trend, we utilize the approach outlined in (Dixon and Pechmann, 2008). To test for whether a trend is negligible, the typical hypothesis test for a regression coefficient is inverted and split instead into two one-sided tests.

$$\begin{aligned} H_{0,a} &: \beta_{v,B} \geq U, \qquad H_{1,a} : \beta_{v,B} < U, \\ H_{0,b} &: \beta_{v,B} \leq -U, \qquad H_{1,b} : \beta_{v,B} > -U. \end{aligned} \tag{4.15}$$

Dixon et al. ((Dixon and Pechmann, 2008)) used the following pair of t-statistics to test for these hypotheses:

$$t_{\text{neutral},a} = \frac{\hat{\beta}_{uv} - (-U)}{s(\beta_{uv})}; \qquad t_{\text{neutral},b} = \frac{U - \hat{\beta}_{uv}}{s(\beta_{uv})}$$

and obtained the corresponding p-values for the probability of the alternative hypothesis by taking the maximum of the p-values associated with the t-statistics $t_{\text{neutral},a}$ and $t_{\text{neutral},b}$, respectively, with significance α and with degrees of freedom $n-2$.

To initialize the iterative search procedure, all individual nodes $u \in 1, ..., n$. We calculate $M_0(u)$ for all $B_0(u) = u$ following the procedures from 4.3.3 at iterative step $k = 0$. Within $M_0(u)$, we calculate each normalized $W_{uv,t}|A_{uv,t}$ by the following equation for all v that are significantly connected to u across all time T:

$$Z_t(u, v) = \frac{W_{uv,t} - \mathbb{E}[W_{uv,t}|A_{uv,t}]}{\mathrm{Var}(W_{uv,t}|A_{uv,t})}$$

where

$$\mathbb{E}[W_{uv,t}|A_{uv,t}] = \frac{\frac{s_{u,t}s_{v,t}}{s_{T,t}}}{\frac{d_{u,t}d_{v,t}}{d_{T,t}}}; \quad \mathrm{Var}(W_{uv,t}|A_{uv,t}) = \left(\frac{\frac{s_{u,t}s_{v,t}}{s_{T,t}}}{\frac{d_{u,t}d_{v,t}}{d_{T,t}}}\right)^2 \kappa_t$$

Next, we find the linear trends of each $Z_t(u, v)$ across time $t = 1, ..., T$ and take the nodes with trends that are either significantly positive or negative. We write $\mathbf{Z}(u, v)$ as the vectorized time series of $Z_t(u, v)$. The trend is calculated as the coefficient with time $t = 1, ..., T$ from ordinary least squares (OLS), between nodes u and v. $\hat{\beta}_{uv}$ is determined to be significantly increasing, decreasing, or stable (neutral) using the method described in the following section 4.3.6, but only using a single node v in place of a set B. If β_{uv} is significant at the α level (in OLS), then denote the nodes v that are significantly connected and increasing or decreasing with initializing node u as v^{**}. We construct an initializing set B_1 with these nodes $\{u, v^{**}\}$ for step $k = 1$.

4.3.8 Iteration and Overlap Filtering Steps

After the procedures for selecting nodes that are both significant in connectivity (Section 4.3.3) and trend $\beta_{v,B}$ depending on the posited direction of trajectory (Section 4.3.6), we derive p-values from the t-statistic of the time-trend . The nodes whose trends are significant after incorporating the FDR correction with significance level α, are retained.

We update the set B_{k+1} with the inclusion of the new nodes v that are both significantly connected to B_k across all time t and have a significant trend according to the trend hypothesis. The procedure is repeated until the set becomes stable such that $B_k = B_{k+1}$ for all candidate sets. In all the applications used in this study, this process takes 3 to 5 iterations.

After stable sets are found from the iteration steps, they are filtered by their Jaccard overlaps (Palowitch et al., 2018; He et al., 2020b). We use an overlap threshold of 0.50 to remove clusters with over 50% overlap; more details on this procedure can be found in prior work (Palowitch et al., 2018). After filtering by Jaccard overlaps, communities of size 3 or less are removed, as dyadic or triadic relationships between nodes may be too localized to be meaningful in a larger scale.

4.3.9 Effect of Normalizing Edges

Modeling network time-series using the weighted configuration model places edge-weights in a relative scale when they are normalized by their expectations and variances, which are functions of global κ_t. Global κ_t is shown to be highly seasonal (fig. 4.2) in the Divvy system in Chicago but less so for the NYC taxicab and Citibike data. The variances in the taxicab data experience a sudden increase in the middle of 2017 and thereafter consistently increase through time. The high seasonality of κ_t in Chicago and the effects of its removal by normalization are apparent in figure 4.4. Scaling edge weights is especially useful in time-series networks where seasonal effects dominate much of the variation (in the Divvy data) or the trend (in NYC taxicab data).

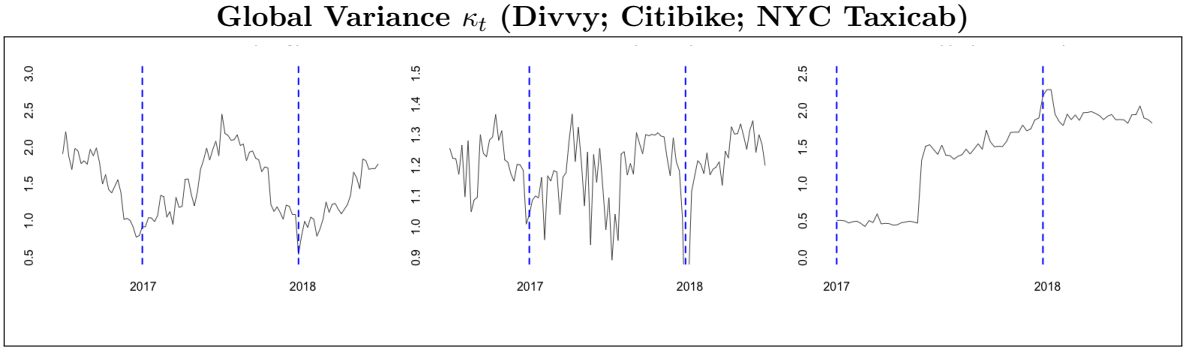


Figure 4.2: Global variance parameter κ_t from 2016 to 2018 for the Divvy system in Chicago (left), the Citibike system in New York City (center), and κ_t for NYC taxicab networks (right) from 2017 to 2018

4.4 Corrections for Forgone Trips Due to Load Imbalance

We use the term load imbalances to refer to the issue of *supply-demand mismatch*,which is a common problem in bikeshare systems where nodes represent stations (which hold bikes) and edges represent the trips between these stations. A station in a high-activity area of a large system is often empty or full during peak hours (Gast et al., 2015; Xie and Wang, 2018; Pendem, 2019). A

full or empty station prevents an otherwise possible trip we thus describe methods (also based on significance testing) of correcting for these empirical inefficiencies in order to assess the true rates of network flow without capacity constraints. We first describe the corrections if there exists load rebalancing data.

We estimate the functionals $\mathbb{P}(\tilde{E}_{u,Y})$ by taking the average rate at which a station is empty i.e. yields no available bikes. $\mathbb{P}(\tilde{E}_{u,Y})$ are calculated as the ratio of the time-intervals that a station is empty to the total intervals during peak-times (i.e. when users could plausibly check out or return bikes). The ratio represents the probability of a station being empty when a user accesses it. A high ratio signifies that the station is usually empty, and so it is more frequently load-imbalanced due to high usage, hence more weight should be proportionally accounted for to estimate the trips that could have been taken if the system was perfectly balanced.

Let $\tilde{E}_{u,Y}$ be the event that a typical trip in year Y from or to station u is foregone owing to load imbalance and let $\mathbb{P}(\tilde{E}_{u,Y})$ be its associated probability. $\mathbb{P}(\tilde{E}_{u,Y})$ is approximated as:

$$\mathbb{P}(\tilde{E}_{u,Y}) \approx \frac{\#\{\text{intervals when } u \text{ is empty in year } Y\}}{\#\{\text{total intervals in station } u \text{ in year } Y\}}. \tag{4.16}$$

Observed demand $W_{uv,t}$ for each edge between stations u, v during time-index t (weeks in this analysis) are then converted to estimated demand $\tilde{W}_{uv,t}$ as follows:

$$\tilde{W}_{uv,t} = W_{uv,t}(1 + \mathbb{P}(\tilde{E}_{u,Y}))(1 + \mathbb{P}(\tilde{E}_{v,Y})), \quad t \in Y.$$

We refer to the time-series of graphs comprised of these demand-corrected (DC) weights as $\{\tilde{G}_t\}_{1 \le t \le T}$. For this study, we assume that this probability is constant over the year. Seasonal effects may be influential in this calculation but will be deferred to future research. We assume that a full station induces a negligible impact on load imbalance compared to empty stations. Each probability is calculated as the proportion of time-intervals that the station is empty. We construct networks of estimated demand to correct for trips that could not have taken place due to full or empty stations and find communities within these networks to more accurately find communities of trip demand in a human mobility network (Faghih-Imani and Eluru, 2015; Liu et al., 2016).

4.4.1 Forgone Trip Corrections Without Rebalancing Data

Though real-time data on station status (e.g. number of open slots) exist and are available online (Divvy, 2019), we do not have access to the historical load rebalancing data and as such we need to estimate the probability of foregone trips. To determine the presence of these forgone trips induced by full or empty stations, we look for anomalous gaps in usage of stations on the days that it is heavily utilized. We refer to these gaps due to forgone trips as *load imbalance*. We describe a simple significance-testing based method that corrects the counts of trips between stations (edge weights) in each graph G_t for week t in each year Y. We have omitted the results of this analysis of the Divvy System in Chicago, though results from this study can be made available on request.

Let $\tilde{E}_{u,Y}$ be the event that a typical trip in year Y from or to station u is foregone owing to load imbalance, we write $\mathbb{P}(\tilde{E}_{u,Y})$ as its associated probability. For this chapter, we assume that this probability is constant over the year. Seasonal effects may be influential in this calculation but will be deferred to future research.

Sums-of-trips $W_{uv,t}$, or *observed demand*, for each edge between stations u, v during time-index t (weeks in this analysis) are then converted to **estimated demand** $\tilde{W}_{uv,t}$ as follows

$$\tilde{W}_{uv,t} = W_{uv,t}(1 + \mathbb{P}(\tilde{E}_{u,Y}))(1 + \mathbb{P}(\tilde{E}_{v,Y})), \quad t \in Y$$

We refer to the time-series of graphs comprised of these demand-corrected weights as $\{\tilde{G}_t\}_{1 \leq t \leq T}$. We now describe how to estimate the functionals $\mathbb{P}(\tilde{E}_{u,Y})$.

4.4.2 Calculating Significant Gaps in Station Activity

A time interval for station u is an interval between any two consecutive events (arrivals or departures). We first formulate a methodology to judge if a time interval is anomalous or not. We call such an unnaturally long time interval a gap. Gaps may occur because of load imbalance or random events not related to load imbalance. We posit that the probability of the occurrence of a foregone trip is:

$$\mathbb{P}(\tilde{E}_{u,Y}) \approx \frac{\#\{\text{gaps in station } u \text{ in year } Y \text{ due to load imbalance}\}}{\#\{\text{intervals between trips in station } u \text{ in year } Y\}}. \tag{4.17}$$

We assume that typical waiting times (in seconds) between consecutive events (start and end of trips) at a station u on day d, $w_{u,d}$ follows an exponential distribution with mean $\delta_{u,d}$ (Gast et al., 2015). Note that the cardinality of waiting times is equivalent to the strengths $S_{u,d}$, or sum-of-trips, of station u on day d subtracted by 1. We filter out the first and last 10% of trips that occurred during day d are censored to filter out the longer gaps during the early and late times of the day, hence only restricting the times s to non-dormant hours, so let $S^*_{u,d} - 1$ represent the number of trips excluding the first and last 10% of trips. We count the number of anomalies *per day* assuming that high-activity stations are rebalancing at least several times a day (Pendem, 2019). To determine anomalies in durations between activity, we first define waiting-times. Let $\theta_{1,u,d} < \theta_{2,u,d} < ... < \theta_{S^*_u,u,d}$ denote the time points of consecutive activity on day d at station u after removing the upper and lower 10% of trip-times.

Let $\mathcal{S}^*_{u,d}$ represent the collection of intervals $\{[\theta_{i,u,d}, \theta_{i-1,u,d}]\}$ and let $w_{i,u,d} = \theta_{i,u,d} - \theta_{i-1,u,d}$ denote the length of these corresponding intervals. We define the sample mean $\bar{\delta}_{u,d}$ as

$$\bar{\delta}_{u,d} = \frac{1}{S^*_{u,d} - 1} \sum_{i=1}^{S^*_{u,d}} w_{i,u,d}.$$

Let $I_{u,d}$ be the number of time-intervals $w_{u,u,d} \in \mathcal{S}^*_{u,d}$ whose lengths are significantly greater than $\delta_{u,d}$ under significance level α after being corrected by the Benjamini-Hochberg false-discovery rate rejection procedure (Benjamini and Hochberg, 1995). This procedure will be described in the later section 4.3.4 and will be used in the community detection algorithm. Precisely:

$$I_{u,d} = \#\{w_{i,u,d} : w_{i,u,d} > \bar{\delta}_{u,d} \text{ at } \alpha, \text{ FDR corrected across } w_{i,u,d} \in \mathcal{S}^*_{u,d}\}.$$

$I_{u,d}$ represents the estimated number of gaps in waiting-times. These values may represent gaps due to either load imbalance or typical events such as a break in usage during lunch, or an

adverse weather event. We assume that these typical events are different from load imbalance. We do not have data on events that could have led to these gaps caused by typical events. However, we can determine a summary measure of the gaps that occurred when the station is operating *in excess*, which we define as the condition when the number of trips is significantly greater than the number of slots in the stations. We can also determine the total sum of the gaps that may be due to random, *typical*, conditions when the station is not operating in excess. We posit that the difference of the gaps under these two conditions provides a reasonable approximation of the gaps owing to load imbalance.

4.4.3 Finding Stations with Excess Demand

We define $C_{u,Y}$ as the carrying capacity, or number of slots, in a station u in year Y. Typically, carrying capacities of stations are updated once per year. If $C_{u,Y}$ of a station (in and outflows) are exceeded significantly at a given day d by the total trips (daily strengths) $S_{u,d}$, then we consider the possibility of a overfilled or empty station may influence the decisions of a potential user. We define *excess demand* $D_{u^*,d}$ in stations u^* where $\{u^* : S_{u^*,d} \geq C_{u^*,Y}\}$ as:

$$D_{u^*,d} = (S_{u^*,d} - C_{u^*,Y}) \sim \text{Poi}(\lambda_d) \tag{4.18}$$

We assume that the counts of excess demand on day d at station u adheres to a Poisson distribution across all stations $u \in [n]$ on day d. Functionals related to the total number of trips between periods of times are conventionally modeled as Poisson (Gast et al., 2015). Let λ_d be the typical network-level excess level of demand in day d and let $\bar{\lambda}_d$ be its sample mean:

$$\bar{\lambda}_d = \frac{1}{n} \sum_{u=1}^{n} D_{u,d}.$$

To determine whether station u is operating *in excess* on a given day d in year Y, we use the Benjamini-Hochberg false-discovery rate correction (section 4.3.4) to find the stations that are significantly over capacity on day d. We evaluate the p-value of excess demand $D_{u,d}$ at station u

by testing every $u \in [n]$ on day d against the sample mean $\bar{\lambda}_d$ under a Poisson distribution under fixed significance α.

We introduce a binary random variable $Q_{u,d}$ to denote if a station is significantly in excess. Let the value of $Q_{u,d} = 1$ if $D_{u,d}$ is judged to be significantly anomalous from $\bar{\lambda}_d$ under significance level α with false discovery rate correction across stations u^* with excess demand above 0, otherwise, let $Q_{u,d} = 0$. Note that $Q_{u,d}$ is zero for all u such that $\{u : S_{u,d} < C_{u,Y}\}$, but it is zero for *some* stations u^* such that $\{u^* : S_{u^*,d} \geq C_{u^*,Y}\}$.

4.4.4 Estimating Foregone Trips

Gaps may be due to typical *baseline* events or to load imbalance. On a given day, a station may be visited above or below its average rate of activity due to chance. However, if the station significantly exceeds demand (number of trips far exceed the number of slots) on such a day, then there is more reason to believe that the gaps in waiting-times between usage are plausibly related to load imbalance.We approximate the gaps using methods described in the previous sections.

Let $\hat{g}^E_{u,Y}$ denote the total approximated number of gaps in activity in station u over year Y on the days d when the station is operating in excess (i.e. $Q_{u,d}$). We assume that the indicator for station u for a gap is independent of the fact that the station is over capacity on day d. The estimated counts of gaps when the station is operating in excess is expressed as:

$$\hat{g}^E_{u,Y} = \sum_{d \in Y} I_{u,d} Q_{u,d}$$

Recall that $1 - Q_{u,d}$ denotes the judgement by the FDR procedure of a non-anomalous demand on day d. Let $\hat{g}^b_{u,Y}$ denote the sum of the number of gaps on days when the excess demand of station u is not significantly anomalous with respect to $\text{Poi}(\bar{\lambda}_d)$. Here $1 - Q_{u,d} = 1$ representative of a typical day with *baseline* anomalies. These counts are estimated as:

$$\hat{g}^b_{u,Y} = \sum_{d \in Y} I_{u,d}(1 - Q_{u,d})$$

Here $\hat{g}^b_{u,Y}$ represents the *natural* number of anomalous gaps from the days not distorted by too much activity in a station that would give rise to full or empty stations. In contrast, $g^E_{u,Y}$ represents an estimate of anomalous intervals (gaps) in stations owing to excess demand. We assume load imbalance can only occur when there is excess demand, and gaps due to excess demand comprise baseline and baseline gaps. We remove the baseline gaps from gaps owing to excess demand by subtracting $\hat{g}^b_{u,Y}$ from $g^E_{u,Y}$ to refine the estimate of gaps induced by load imbalance. Because load imbalance can only decrease the efficiency of the system by reducing the number of trips, the demand-correction probability can only be increased and the numerator of (4.17) is:

$$\#\{\text{gaps due to load imbalance in station } u \text{ in year } Y\} \approx \left(\hat{g}^E_{u,Y} - \hat{g}^b_{u,Y}\right)^+ \tag{4.19}$$

The probability of a forgone trip (4.17) can be estimated by

$$\mathbb{P}(\tilde{E}_{u,Y}) \approx \frac{\left(\hat{g}^E_{u,Y} - \hat{g}^b_{u,Y}\right)^+}{\sum_{d\in Y}(S^*_{u,d} - 1)} \tag{4.20}$$

where the denominator, which represents the total number of time-intervals in all days across year Y, can be represented by the sum of trips (daily strengths excluding first and last 10% of trips) of station u in each day d. We use these probabilities to construct a demand-corrected time-series of graphs $\{\tilde{G}_t\}_{1\le t\le T}$ and find communities in these networks in addition to the uncorrected graphs.

4.5 Results

We report results for a range of values for tuning parameters α and U for observed demand $\{G_t\}_{1\le t\le T}$. In the Divvy Network, we fix α at 0.05 and $U = 0.007$ as well as 0.009 because these settings capture clusters of moderate sizes across all trend categories and also show distinct geographical divisions. Under these tuning parameters, we find five clusters with decreasing connectivities over time and five clusters with increasing connectivities. We find only one cluster with a stable trend at the 0.05 significance level.

There is a stark division in trends between the northern and southern parts of the city (fig. 4.3). At the 5% significance level, clusters with significantly decreasing trends are mostly found in the southern and western parts of the city, while clusters with significantly increasing trends are

mostly found in the northern and central parts of the city. Interestingly, the decreasing clusters map to a nearly concentric outer ring around the central part of the city, while the increasing clusters stretch from the Loop northwards along the shore of Lake Michigan. One stable cluster is located in the Loop.

It is useful to focus on one community to illustrate the effect of edge normalization (see section 4.3.3). In figure 4.4, while the raw edge weights show a stable trend, the normalization $\mathbf{Z}(B)$ shows an increasing trend. Thus, the collection of five stations in the Lincoln Park neighborhood in Chicago is classified as a cluster with an increasing time trend rather than a stable one.

The geographical domain that the taxicab network covers is much larger than the bikeshare network, which only spans Manhattan and Brooklyn. In two settings of U, clusters are decreasing in connectivity across much of the Bronx, Queens, and much of Brooklyn. Clusters are consistently increasing in eastern parts of Queens. One cluster appears to consistently link Staten Island to southern Brooklyn for both values of U. Clusters are stable around the denser parts of the city, as is the case in Upper Manhattan when U is 0.01 and in Upper and Lower Manhattan, Central Brooklyn, and Astoria in Queens when U is 0.02.

4.5.1 Effect of Demand Correction

We apply the intertemporal community detection algorithm to the demand-corrected (DC) time-series networks $\{\tilde{G}_t\}_{1\leq t\leq T}$ with weights $\tilde{W}_{uv,t}$ in the Citibike system. We use the same significance $\alpha = 0.05$ and set barrier U to 0.007 and 0.009 as in observed trip networks in NYC and the Divvy system in Chicago. The obtained communities retain similar geographical characteristics as those in uncorrected graphs, but with some key differences.

When U is set at 0.007, the decreasing and increasing clusters in the demand-corrected networks are localized in approximately similar geographical regions as in non-corrected networks. Increasing clusters are mostly located in Upper and Lower Manhattan as well as Southern Brooklyn. Decreasing clusters are present in some small areas throughout Manhattan but pervasively cover swathes of northern Brooklyn around the Williamsburg region. Stable clusters mostly span Midtown Manhattan but also extend to northern Manhattan and parts of Brooklyn.

When U is increased to 0.009, the increasing and decreasing clusters shrink in size and number and the stable clusters expand. Increasing clusters are more visibly located in Upper and Lower

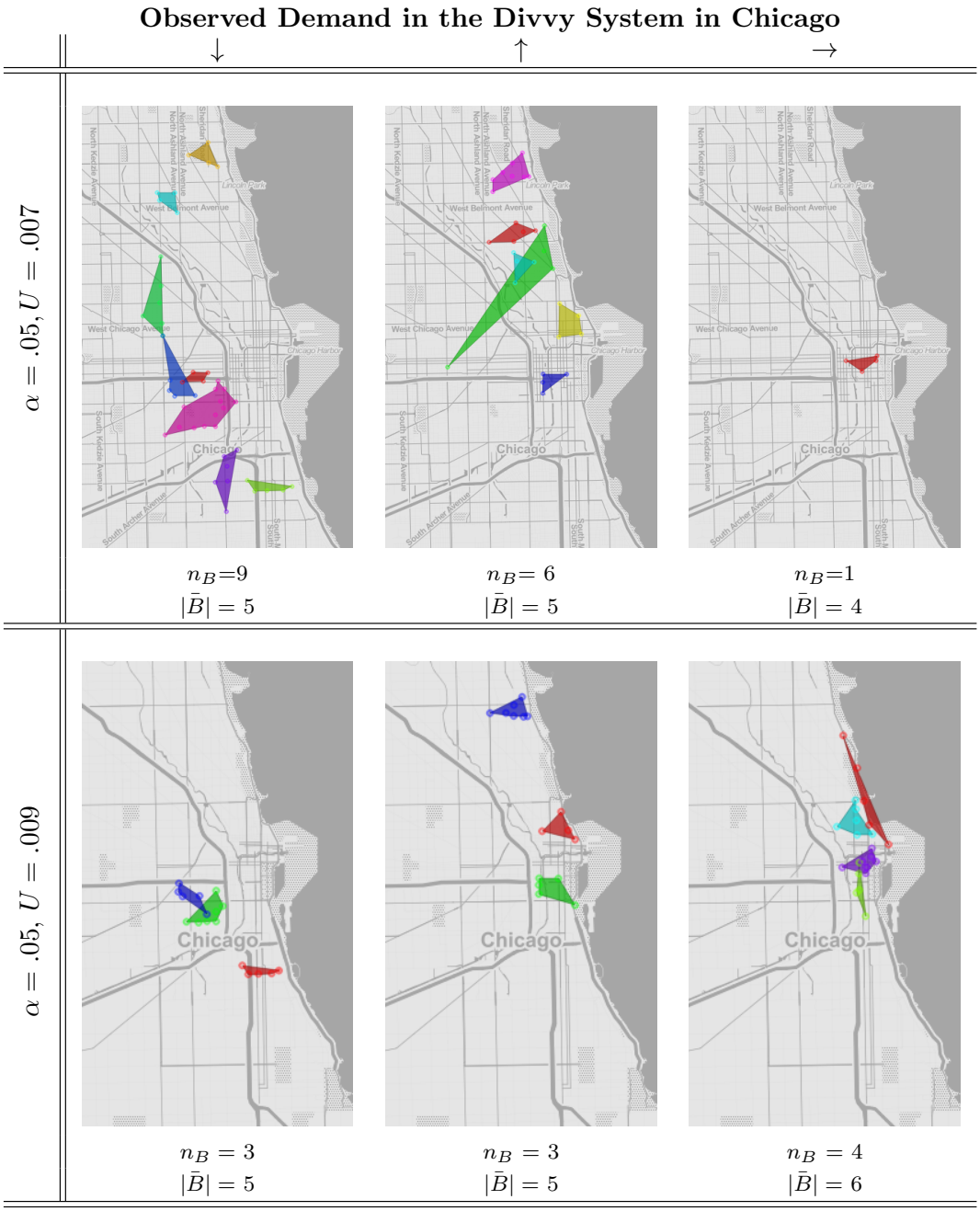
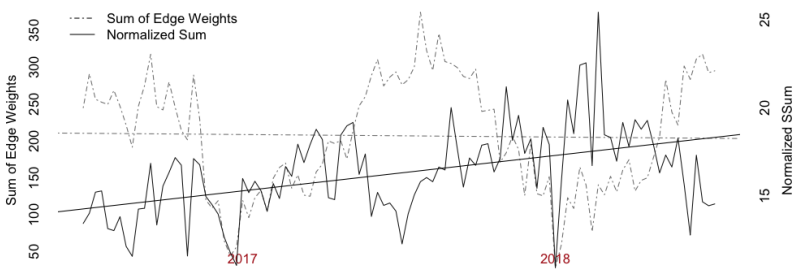


Figure 4.3: Intertemporal communities of increasing or decreasing trends amongst Divvy stations in 2016-2018 under varying significance levels and bounding parameters U using the network time-series $\{G_t\}$ uncorrected for load-imbalance. n_B represents the number of found communities and $|\bar{B}|$ represent the mean size of communities.



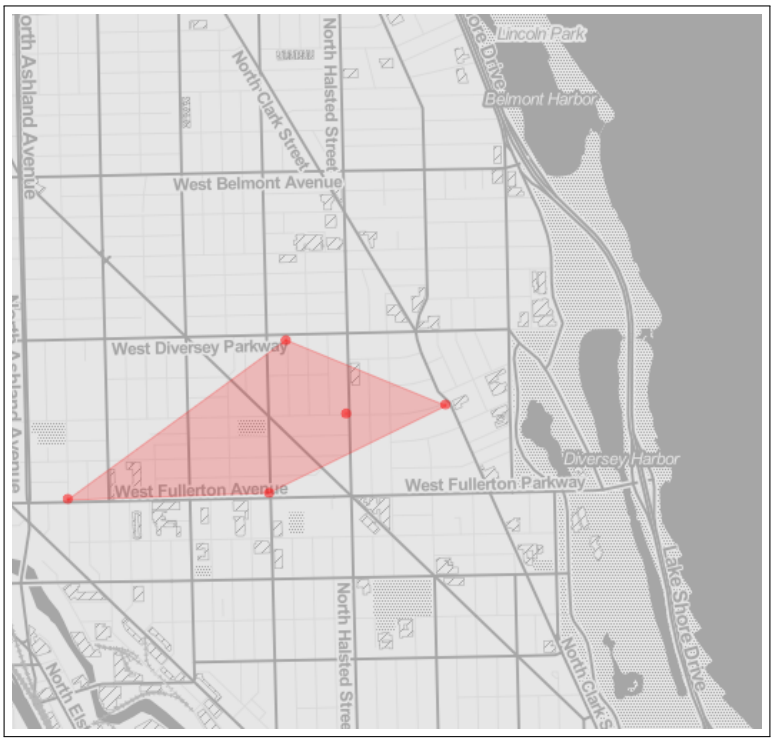


Figure 4.4: *top*: Total trips in a community in networks G_t with increasing normalized connectivity over time comprising 5 stations around the Lincoln Park Neighborhood in Chicago. *bottom:* Map of stations in B.

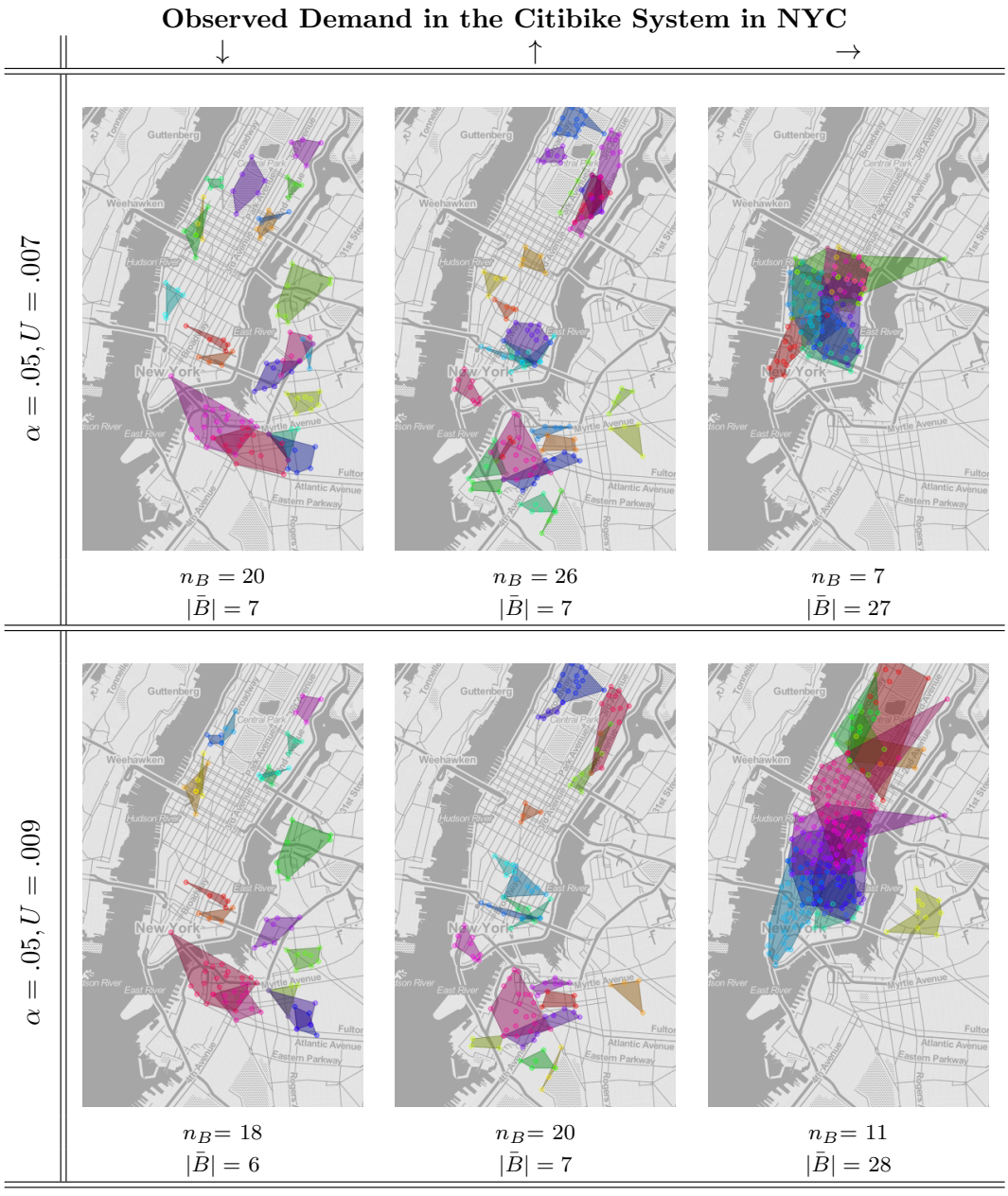


Figure 4.5: Intertemporal Communities of increasing (↑), decreasing (↓), and stable (→) trends amongst stations in years 2016-2018 under varying significance levels and bounding parameters U in the uncorrected networks G_t. n_B represents the number of found communities and $|\bar{B}|$ represents the mean size of communities.

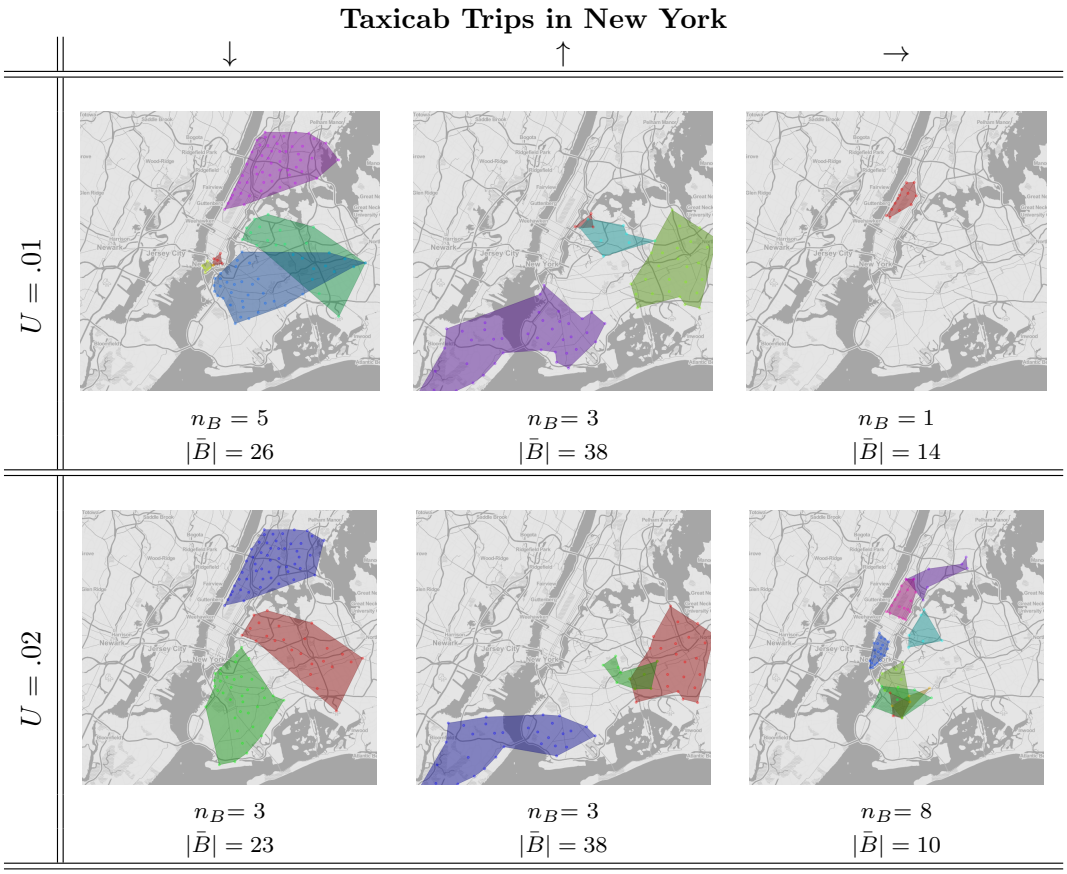


Figure 4.6: Intertemporal Communities of increasing, decreasing, and stable trends in taxicab networks amongst zones in years 2017-2018 in New York City under varying significance levels and bounding parameters U. n_B represents the number of found communities and $|\bar{B}|$ represents the mean size of communities rounded to the nearest integer.

Manhattan (similar to the clusters in the graphs of observed demand) at the higher threshold. Decreasing clusters are interspersed throughout the city but large coherent areas are more clearly located around northern Brooklyn, also as in the observed graphs G_t. The stable graphs, however, are much larger and cover much more ground in Lower Manhattan (fig. 4.7) .

4.6 Discussion

In the Citibike, Divvy, and NYC taxicab systems, we observe a trade-off between increasing or decreasing clusters and stable clusters depending on the choice of U. If U is larger, then there is "more room" for a trend to be classified as stable, but less so for increasing or decreasing trends. Discovery of more increasing and decreasing clusters when U is increased suggests that these clusters are increasing or decreasing in connectivity at different rates from the other clusters. When U is large, increasing and decreasing clusters vanish but more stable clusters persist.

The interaction between α and U is not entirely linear or monotonic. Though a decrease in α may correspond to an increase in U, a lower α implies that the nodes are more connected at each time-instance and does not necessarily mean that the trend is higher. Figures 4.5 and 4.7 show that in both G_t and $\tilde{G}_t$, clusters appear as U becomes larger and α stays the same. Such behavior may be attributed to FDR correction. A lower barrier U may yield more significantly connected nodes but with weaker trends. The sensitivity of community detection to the choice of parameter is an important issue (Austwick et al., 2013). We compare the extracted communities under different tuning parameters U and α.

In results from the observed network G_t in Chicago, the choices of α and U produce generally similar results over a range of values (fig. 4.3). Shifting U from .007 to .009 induces discovery of more increasing and decreasing clusters, but the bound is too tight for any significant sets to be found under the hypothesis tests in (4.15).

Our analysis is exploratory in nature and only summarizes the trajectories of network structures in time but not their underlying causes. While this work is focused on methodological aspects of temporal community detection, results suggest that it might be useful to think about the causal mechanisms that underlie the different types of clusters in the bikeshare networks. The geographical patterns of the cluster map to different neighborhood characteristics.

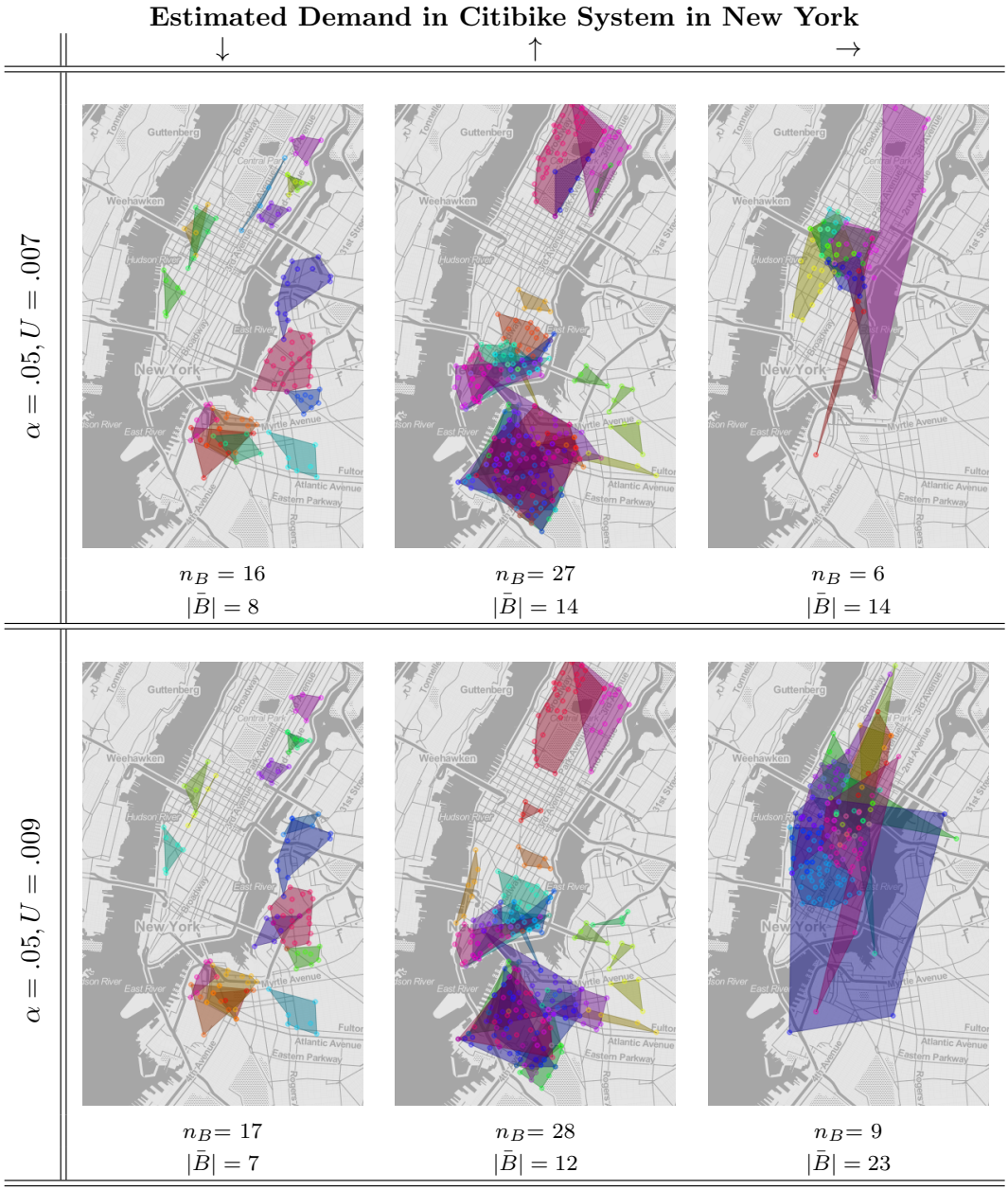


Figure 4.7: Intertemporal Communities of increasing, decreasing, and neutral trends amongst Citibike stations in years 2016-2018 in New York City under varying significance levels and bounding parameters U in the demand-corrected networks $\tilde{G}_t$. n_B represents the number of found communities and $|\bar{B}|$ represents the mean size of communities rounded to the nearest integer.

At the same fixed parameters for U and α, clusters in NYC are more numerous and less geographically spread out than Chicago, possibly because the city is much denser and more populous. Moreover, the seasons are milder, which induces less variation in trends. Figure 4.2 shows that the global variance parameter κ_t of the Divvy system is highly seasonal, unlike that of the Citibike system. Several areas appear to be persistently decreasing in both demand-corrected and observed networks.

The edge-normalizing step of the community detection algorithm (section 4.3.9) makes such station-specific adjustments affect the whole network, thereby affecting the entire system. Regardless, similarities persist in clusters in both DC and uncorrected graphs. The demand corrected (DC) networks $\tilde{G}_t$ when U is 0.007 and 0.009 yield similar decreasing clusters to those of the uncorrected networks G_t. Demand-adjustment makes a considerable difference in some clusters in the Citibike system. Adjusting for demand-correction thus reveals stronger, more cohesive increasing trends within the ridership and suggests that observed trips do not adequately capture the latent increasing signals that are distorted by load imbalances from empty stations.

Because the proposed method is for exploratory purposes, these summarizing claims should be verified in a more rigorous way in future research. Furthermore, the choice of U varies by application. We use an ad-hoc scheme to select U whose resultant neutral clusters yield approximately the same amount of nodes as the increasing and decreasing clusters combined.

However, because most of the results we present include two different values of U in order to show the differences in results due to adjusting the parameters , the results in this study may not strictly adhere to this criteria. However, results from Fig. 4.3, Fig. 4.5 and 4.7 all approximately follow ths heuristic when $U = .009$, though they may have different α's. Different applications of intertemporal community detection may call for different criteria for tuning parameters. For example, setting U to be small so as to not allow discovery of any neutral clusters (i.e. Fig. 4.3, $U = .007$) may also be a suitable option. In future work, more principled approaches for setting tuning parameters utilizing cross-validations may be investigated.

4.6.1 Future Work

In both Chicago and New York City, there may be several explanations for the underlying signals that cause the clusters to decrease in connectivity. Further work may examine what these

signals are and how these signals may function. One explanation may be that decreasing trends are symptoms of displacement, destabilizing steady ridership among long-term inhabitants in gentrifying neighborhoods. Another may be differential rates of attention given to load rebalancing in stations in different neighborhoods with varying resources. Causal analysis of these phenomena are outside the scope of this study, but our exploratory results are useful in initializing conversations about changes in mobility patterns within and between neighborhoods. Future work may analyze the relationship between the discovered communities and factors such as new construction, bike lanes, weather, incomes, and demographic characteristics.

The methods devised in this study can be applied to a variety of data in network time-series format, particularly human mobility networks. The method can be applied to bikeshare networks in other cities, or may be applied to other networks of transportation in urban systems. Future work may elaborate on the theoretical properties of intertemporal community detection. The null model described in section 4.3.1 may also have further use in statistical inference or in forecasting future patterns. Another extension would be to account directly for the spatiotemporal aspects of trips in the methodology.

Our work currently relies on historic station inventory data for the analysis of the Citibike system. We do not have access to historical inventory data for Chicago and thus are not able to estimate demand. Though similarities between corrected and non-corrected networks in the NYC bikeshare system shows that there may be some use in using only non-corrected data in Chicago, there are limitations in drawing conclusions for demarcations of functional mobility zones using only observed demand. We propose a method in 4.4.1, but further estimation of demand without historical station inventory data should be explored in future work in conjunction with community detection.

We proposed a novel method to cluster networks representing bikeshare systems that vary across time. Our community detection method combines usage of a configuration null model with a trend model to describe the expected trajectory of the graph evolutions. We use a significance-testing methodology to assess whether nodes are anomalously connected to each other within and across time-periods. By using the proposed method, we are able to filter some of the system-wide seasonal effects and map geographically coherent communities of latent human mobility signals in the bikeshare stations in Chicago and New York and the taxicab network in New York. The

methods used in this chapter may be applied to other situations where it is important to study the evolution of structures within networks.

CHAPTER 5

Bimodules Clustering for Bipartite Correlation Networks

With the development of high throughput data in fields such as genomics, neuroscience, and atmospheric science, researchers often need to compare two or more data sets derived from a common set of samples[1]. In most cases, different technologies measure different features and capture different information about the samples. While data arising from different settings may be separately analyzed, additional and potentially fundamental insights can sometimes be gained from the joint (or integrated) analysis of the data sets. Multi-modal analysis has received considerable attention in present literature (Lahat et al., 2015; Meng et al., 2016; Tini et al., 2019; Pucher et al., 2019; McCabe et al., 2019) .

We develop a new method to search for correlated groups of bipartite variables. We first define the groups of variables (called bimodules) that need to be discovered. A bimodule is a minimal group of variables from the two data types such that variables of the two types within this group are correlated with each other in aggregate, but no variable within this group is strongly correlated with variables of the other type outside the group.

Firstly, even though there are many groups of variables compared to the total variable dimension, BSP adaptively searches this space by iteratively updating the variable groups, guided by the results of many hypothesis-tests at each step. We use analytical approximations to quickly compute the p-values for these tests. Secondly, BSP directly analyzes the primary data rather than only the correlation matrices. Therefore, the method is statistically motivated and "borrows strength" from the interactions between variables of the same type during the search algorithm. Accounting for these interactions in the algorithm thus make the method implictly network-analytical, even if the inputs of the algorithm are not necessarily adjacency matrices, but directly the observations. Simulation studies suggest that false discoveries under BSP are controlled.

[1]This chapter is adapted from a manuscript from 2020 (Dewaskar et al., 2020), this was joint work with Miheer Dewaskar (primary author), Andrew Nobel, and Michael Love

BSP relies on permutation-based p-values for test statistics equal to sums of squared cross-correlations. These p-values are approximated using tail probabilities of gamma distributions that are fit using the estimates of the permutation moments' test statistic. BSP moment estimates depend on the eigenvalues of the intra-correlation matrices between two data types, and as a result the significance of observed cross-correlations accounts for the correlations within each data type.

5.1 Layout and Contributions

We first give an abridged background on the theory and methodological details of the search procedure, then discuss the novel application of this method to climate data. The contributions to this method is in its novel application to climactic time-series data. Furthermore, embedded in this application is the novel conception of gridded temperature and precipitation time-series as bipartite networks. Because the overall sample size of these climate networks is much smaller than in genomic data due to aggregation of resolution, more computationally intensive approaches may be used for identifying optimal parameters such as the false discovery rate based on edge-error estimates. We are able to repeat computations described in section to make more robust inferences about optimal choices of significance parameters.

This chapter will proceed as follows: we first describe prior approaches that are similar in motivation or application to BSP in the following section 5.2. We then set up the notation and establish the setting for the data in section 5.3. We then define the theoretical bimodule object in section 5.3.1 and its empirical (sample) counterpart in section 5.4 and its associated search procedure. Within this section, we start by describing the distribution for the null model, then describing the iterative hypothesis test procedure and its associated algorithm, then detailing the initialization and selection of tuning parameter α. Following this, we move onto the application of the method to temperature and precipitation data in North America procured from the Climactic Research Unit (CRU) in Section 5.5. We first describe the data and preprocessing in section 5.5.1, then the method application and results in Section 5.5.2. Finally, we describe the results on the genomic analysis in Section 5.6.

5.2 Prior Work on Bimodules

Since bimodules are defined in terms of cross-correlations, it is natural to investigate them in the context of the bipartite *cross-correlation* network. Such a network is formed by connecting pairs of cross-correlated features with an edge having a weight equal to the square of their (sample or population) correlation. CONDOR (Platig et al., 2016) identifies bimodules by applying community detection to an unweighted bipartite graph obtained by thresholding the sample cross-correlations. One could, in principle, extend this approach by leveraging other community detected methods (Beckett, 2016; Barber, 2007; Liu and Murata, 2010; Costa and Hansen, 2014; Pesantez-Cabrera and Kalyanaraman, 2016) for weighted and unweighted bipartite networks.

The approach taken here is network based, but differs from community-detection based approaches such as CONDOR. While stable population bimodules can be defined in terms of the population cross-correlation network, the sample cross-correlation network is not a sufficient statistic for stable sample bimodules, which depend on (and account for) intra-correlations between features of the same type. (Huang et al., 2009) also identify groups of associated genes and SNPs by adapting bipartite clique mining, however they work with a tri-partite network derived from progeny strain data. We conduct a comparative study with CONDOR as a competing method described in Section 5.6.1. Other approaches such as sparse Canonical Correlation Analysis (sCCA) as proposed by (Parkhomenko et al., 2009) and Group eQTL (GeQTL) (Cheng et al., 2015) are similar in principle to the proposed method. However, these approachs require pre-specifying the number of clusters and require that every feature be a member of some cluster and also do not distinguish between inter- and intra-correlations, and moreover use different clustering dynamics from iterative testing and thus yield results that are inherently different.

5.3 Notation and Setup

Suppose there are two high-dimensional datasets measuring information on the same n individuals. The measurements of the first type are represented by $n \times p$ matrix $\mathbf{X}$ with n rows (observations) and p columns (samples). Those of the second type are represented by a $n \times q$ matrix $\mathbf{Y}$. In both $\mathbf{X}$ and $\mathbf{Y}$, the ith row of the matrices measuring the (same) ith individual, and the columns corresponding to measured variables. We denote the indices of variables of the two

data types by S corresponding to $\mathbf{X}$ and T corresponding to $\mathbf{Y}$:

$$S = \{s_1, s_2, \ldots, s_p\}, \quad T = \{t_1, t_2, \ldots, t_q\}$$

We assume that the rows of of the joint matrix $[\mathbf{X}, \mathbf{Y}]$ are independent copies of a jointly random vector

$$(\mathbf{X}, \mathbf{Y}) = (X_{s_1}, \ldots, X_{s_p}, Y_{t_1}, \ldots Y_{t_q}) \in \mathbb{R}^{p+q}.$$

We assume spherical symmetry in one of $\mathbf{X}$ or $\mathbf{Y}$. Such an assumption is useful for fast approximations of p-value computations involved in the hypothesis tests. These approximations have been show to work well for gene-expression data.

5.3.1 Bimodules

For each $s \in S$, let $\mathbf{X}_s$ denote the column of $\mathbf{X}$ corresponding to the variable s, and for each $t \in T$ define $\mathbf{Y}_t$ similarly. For any $s \in S$ and $t \in T$ let $\rho(s,t)$ denotes the unknown population correlation between the random variables X_s and Y_t, and let $r(s,t)$ denote the observed sample correlation between the columns $\mathbf{X}_s$ and $\mathbf{Y}_t$. Finally if $A \subseteq S$ and $B \subseteq T$, let

$$\rho^2(A,B) \doteq \sum_{s \in A, t \in B} \rho^2(s,t), \text{ and}$$
$$r^2(A,B) \doteq \sum_{s \in A, t \in B} r^2(s,t).$$

For singleton sets, we omit the surrounding brackets. Hence for $s \in S$ we write $\rho^2(s,B)$ instead of $\rho^2(\{s\},B)$.

The broad aim is to find pairs of sets (A,B), where $A \subseteq S$ and $B \subseteq T$, so that variables in A and B are all correlated with each other, but with no other variables. Specifically, elements in sets A and B are correlated with each other in an aggregate sense, but no variable in the remainder set $T \setminus B$ is correlated with those in A, and similarly no variable in $S \setminus A$ is correlated with those in B. We will call such a pair (A,B) a *bimodule*. Rigorously, we define a theoretical (population) bimodule as:

Definition 5.1. (A,B) of non-empty sets $A \subseteq S$ and $B \subseteq T$ is a *stable population bimodule* if

1. $A = \{s \in S \mid \rho^2(s, B) > 0\}$ and

2. $B = \{t \in T \mid \rho^2(A, t) > 0\}$.

A represents exactly the set of features in S that are correlated in aggregate with the features in B, while B is exactly the set of features in T that are correlated in aggregate with the features in A. For the rest of the chapter, we refer to For theoretical purposes, we model bimodules in the context of the population network of cross-correlations.

Definition 5.2. The *population cross-correlation network* G_p is the weighted bipartite network with vertex set $S \cup T$, edge set $E_p = \{(s, t) \in S \times T \mid \rho(s, t) \neq 0\}$, and weights $\rho(s, t)$ between -1 and 1 .

The following lemma shows that bimodules are closely related to the connected components of G_p.

Lemma 1. *A pair (A, B) of non empty sets with $A \subseteq S$ and $B \subseteq T$ is a population bimodule if and only if $A \cup B$ is a union of non-trivial connected components of G_p.*

Proof. For any subsets $F \subseteq S$ and $G \subseteq T$, we note that $\rho^2(F, G) > 0$ if and only if some pair $(s, t) \in F \times G$ has $\rho(s, t) \neq 0$. Therefore, condition 1 of definition 5.1 says that every $s \in A$ has at least one neighbor in B under G_p, and A is exactly the set of all such neighbors of B.

Similarly condition 2 of definition 5.1 states that every $t \in B$ has at least one neighbor in B under G_p, and A is exactly the set of all such neighbors. Both these conditions are satisfied only when $A \cup B$ is the union of some non-trivial connected components of G_p. □

As the lemma shows, stable population bimodules depend only on the edges of G_p; they do not depend on the edge weights, or on correlations between features of the same type. As we will see below, the situation for sample bimodules is substantially different.

5.4 Sample Bimodules and Search Procedure

We define a *sample bimodule* as an estimated bimodule that is inferred from observed data. In practice, the population cross-correlation matrix $(\rho(s, t))_{s \in S, t \in T}$ is unknown, We use the sample cross-correlation matrix $(r(s, t))_{s \in S, t \in T}$. Since sample correlations are almost always non-zero, for

$s \in S$ and $B \subseteq T$, the condition $\rho^2(s, B) > 0$ there must be replaced with $r^2(s, B) > \tau_{s,B}$, for some threshold $\tau_{s,B} > 0$. Analogously, for $A \subseteq S$ and $t \in T$, the condition $\rho^2(A, t) > 0$ must be replaced with $r^2(A, t) > \tau_{A,s}$, for some $\tau_{A,s} > 0$. We will choose the thresholds

$$\{\tau_{s,B}\}_{s \in S, B \subseteq T} \cup \{\tau_{A,t}\}_{A \subseteq S, t \in T}$$

using principles from multiple hypothesis testing.

We now define the null distribution and the p-values for the iterative hypothesis testing scheme used to cluster bimodules.

Definition 5.3. For a set of given observed data matrices $[\mathbf{X}, \mathbf{Y}]$ with dimensions $n \times p$ and $n \times q$ respectively, and for permutation matrices $P_1, P_2 \in \{0, 1\}^{n \times n}$ chosen independently and uniformly at random, the *permutation null distribution* is the distribution of the data matrix

$$[\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}] \doteq [P_1\mathbf{X}, P_2\mathbf{Y}]$$

The permutation null distribution is obtained by randomly permuting the rows of $\mathbf{X}$, then independently doing the same kind of reordering for rows of $\mathbf{Y}$. The permutation distribution depends on the observed data matrix $[\mathbf{X}, \mathbf{Y}]$. Such an operation preserves the sample-correlation between any two variables within S or within T, but it nullifies the sample-correlation between any variable from S and any variable from T (i.e. makes to zero). The for $s \in S$ and $t \in T$, let $\tilde{r}(s, t)$ denote the sample-correlation within the vectors $\tilde{\mathbf{X}}_s$ and $\tilde{\mathbf{Y}}_t$. The latter statement is justified by a lemma from (Zhou et al., 2013). We first define the use this permutation null distribution to define p-values.

Lemma 2. *Suppose the data matrix* $[\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}]$ *has the permutation null distribution obtained from the matrix* $[\mathbf{X}, \mathbf{Y}]$. *Then for any* $s \in S$ *and* $t \in T$, $\mathbb{E}\tilde{r}(s, t) = 0$, *where the expectation is taken with respect to the permutation null distribution.*

Definition 5.4. Suppose the data matrix $[\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}]$ has the permutation null distribution obtained from the observed data matrix $[\mathbf{X}, \mathbf{Y}]$. For $A \subseteq S$ and $B \subseteq T$ define

$$p(A, B) \doteq \mathbb{P}(\tilde{r}^2(A, B) \geq r^2(A, B)) \tag{5.1}$$

where $\tilde{r}^2(A,B) \doteq \sum_{s\in A, t\in B} \tilde{r}^2(s,t)$, and the probability is computed over the permutation null distribution.

The p-value $p(A,B)$ is the probability that aggregate cross-correlation from variables within A and B under the permutation null distribution, $\tilde{r}^2(A,B)$, exceeds its observed value $r^2(A,B)$. A low value of $p(A,B)$ provides evidence in favor of the hypothesis that $\rho^2(A,B) \neq 0$. Since the permutation distribution preserves the correlations within variables from A and within variables from B, $p(A,B)$ accounts for the presence of these correlations while assessing the significance of $r^2(A,B)$. The p-value can be computed by a Monte Carlo simulation draswing from several random permutations of the data matrix.

We now define sample bimodules wherein conditions on the population correlation ρ are replaced with the corresponding hypothesis test using aforementioned p-values, carried out using the Benjamini-Yekutieli multiple testing procedure.

Definition 5.5. (Sample Bimodule) Fix an $\alpha \in (0,1)$. A pair (A,B) of non-empty subsets $A \subseteq S$ and $B \subseteq T$ is a *sample bimodule* at level α if

1. $A = \{s \in S \mid p(s,B) \leq \tau_\alpha(\boldsymbol{p}_{\cdot,B})\}$, and
2. $B = \{t \in T \mid p(A,t) \leq \tau_\alpha(\boldsymbol{p}_{A,\cdot})\}$,

where $\boldsymbol{p}_{\cdot,B} = (p(s,B))_{s\in S}$ and $\boldsymbol{p}_{A,\cdot} = (p(A,t))_{t\in T}$.

We use the following procedure to find empirical bimodules.

Initialize: Select a singleton set $A_0 = \{s\} \subseteq S$ and let $B_0 = \emptyset$.

Repeat for $k = 1, \ldots, k_{max}$:

- For each $t \in T$ compute the p-value $p(A_{k-1}, t)$ and let $\mathbf{p} \leftarrow (p(A_{k-1},t))_{t\in T}$.
- Let $B_k = \{t \in T \mid p(A_{k-1},t) \leq \tau_\alpha(\mathbf{p})\}$ be the set of $t \in T$ rejected by the Benjamini-Yekutieli procedure.
- For each $s \in S$ compute the p-value $p(s, B_k)$ and let $\mathbf{p} \leftarrow (p(s,B_k))_{s\in S}$.
- Let $A_k = \{s \in S \mid p(s,B_k) \leq \tau_\alpha(\mathbf{p})\}$ be the set of $s \in S$ rejected by the Benjamini-Yekutieli procedure.

– Stop if $(A_k, B_k) = (A_{k-1}, B_{k-1})$.

Output: (A_k, B_k) if both sets are non-empty and $(A_k, B_k) = (A_{k-1}, B_{k-1})$.

If BSP terminates at a non-empty fixed point then its output is a stable bimodule at level α. However, BSP is not guaranteed to terminate in a finite number of steps: there may be a convergent set or a cycling of results as the procedure operates in a deterministic manner. As such, we stop the iterative search after a fixed number of steps, determined by parameter k_{max} in algorithm.

5.4.1 Initialization

We initialize the BSP with each singleton pair $(s, \emptyset)$ for $s \in S$, and each singleton pair $(\emptyset, t)$ for $t \in T$.

The constant $\alpha \in (0, 1)$ is the only free parameter of BSP. While α controls the false discovery rate at each step of the search procedure, this does not guarantee control of the false discovery rate of the stable bimodules, or the false associations (i.e. (s, t) such that $\rho(s, t) = 0$) within the stable bimodules. In general, BSP will find fewer and smaller bimodules when α is small, and find more numerous and larger bimodules when α is large. In practice, we employ a permutation based procedure to select α from a fixed grid of values.

Simulations and theoretical calculations suggest that singleton bimodules at a given level $\alpha \in (0, 1)$ can occur even in completely random data if $|S|$ and $|T|$ are large enough. We eliminate bimodules that fail to be significant – bimodules (A, B) with $p(A, B) > \frac{\alpha}{|S||T|}$. The latter threshold is chosen using Bonferroni correction over all pairs in $S \times T$ to minimize the chance that singleton bimodules are detected in completely random data.

The BSP search procedure may find the same bimodule starting from multiple initializations; we deem these bimodules as equivalent. When there are many bimodules with substantial overlap, we assess the *effective* number of distinct bimodules and select this number of representative bimodules for subsequent analysis.

5.4.2 Choice of α

To select the false discovery parameter α, we estimate the fraction of erroneous essential edges among bimodules at level α. This edge error estimate is calculated by considering the fraction of

essential edges from bimodules that are spurious, when BSP(α) is run on a dataset in which half of the variables of each type are permuted. We describe this half-permuted dataset in the next section.

Comparing results between the original and permuted data allows us to empirically assess the false discoveries by BSP when there are no cross-correlations between variable sets S and T. However, the associations between at least some variables from S and T (in fact, these are the ones that we want to find), and need an estimate on the proportion of false discoveries under such conditions. For this, starting with our original data $D = (\mathbf{X}, \mathbf{Y}, \mathbb{C})$, we generate a *half-permuted* dataset as follows:

1. Randomly select half the features, $\hat{S} \subseteq S$ and $\hat{T} \subseteq T$, from each data type.

2. Randomly permute the rows of the submatrix of $\mathbf{X}$ that corresponding to the columns $\hat{S}$, and call the resulting matrix $\tilde{\mathbf{X}}$. In other words, the rows corresponding to the columns $S \setminus \hat{S}$ are the same in $\mathbf{X}$ and $\tilde{\mathbf{X}}$, and the rows corresponding to features in $\hat{S}$ have been permuted $\mathbf{X}_{\hat{S}} = P_1 \mathbf{X}_{\hat{S}}$.

We call this the half-permuted data $H_S \doteq (\tilde{\mathbf{X}}, \mathbf{Y}, \mathbb{C})$. Note that the "half-permutation" in step 2 removes the effect of inter-correlation between (variables in) S_p and T, and the intra-correlations between S_p and $S \setminus S_p$. However, intra-correlations within S_p, $S \setminus S_p$ and T are retained.

5.4.3 False discovery rate based on half-permutation

Let $\mathcal{B} = \{(A_1, B_1), (A_2, B_2) \ldots (A_K, B_K)\}$ be a collection of bimodules obtained from BSP(α) on the permuted dataset after filtering for overlaps, and suppose $S_p \subseteq S$ and $T_p \subseteq T$ are subsets that have been permuted (S_p or T_p may be empty sets). Using this we define the edge-error estimate

$$\text{edge-error}(\mathcal{B}) = \frac{1}{K} \sum_{i=1}^{K} \frac{|\text{essential-edges}(A_i, B_i) \cap S_p \times T \cup S \times T_p|}{|\text{essential-edges}(A_i, B_i|} \tag{5.2}$$

The above edge-error estimate should be used as follows. First, generate multiple instances of the half-permuted dataset. Next choose a grid, for example $\{0.01, 0.02, ..., 0.05\}$, and for each α from the grid, run BSP(α) over the each of the half-permuted datasets and calculate the average edge-error over each of these bimodules. Then we choose an α from the grid that has edge-error

smaller than a pre-specified threshold like 0.05. Usually smaller values of α tend to have smaller edge error. Hence we choose largest value of α with a small enough edge-error. The above edge-error estimate may be variable. The false discovery rate estimates are used to select the value of $\alpha \in (0, 1)$ used for BSP.

5.5 Application to Clustering of Temperature and Precipitation in North America

The relationship between temperature and precipitation over North America has been well documented (Madden and Williams, 1978; Berg et al., 2015; Adler et al., 2008; Livneh and Hoerling, 2016; Hao et al., 2018) and is of agricultural importance. We applied BSP to find pairs of geographic regions such that summer temperature in the first region is significantly correlated with summer precipitation in the second region one year later. We will refer to such region pairs as T-P bimodules. T-P bimodules reflect mesoscale analysis of region specific climactic patterns, which can be useful for predicting impact of climactic changes on practical outcomes like agricultural output.

5.5.1 Data Description and Processing

The Climactic Research Unit (CRU TS version 4.01) data (Harris et al., 2014) contains daily gridded global measurements of temperature (T) and precipitation (P) levels over land at a resolution of $.5^o \times .5^o$ (360 pixels by 720 pixels) from 1901 to 2016. We reduced the resolution of the data to $2.5^o \times 2.5^o$ (72 by 144 pixels) by summing over an aggregating grid of the reduced resolution, and restricted the resulting data to 427 pixels that corresponded to the latitude-longitude pairs within North America. For each available year and each pixel/location we summed temperature (T) and precipitation (P) over the Summer months of June, July, and August. Each of the resulting time series was centered and scaled to have zero mean and unit variance. The data matrix $\mathbb{X}$, reflecting temperature, has 115 rows containing the annual summer-aggregated temperatures from 1901 to 2015 for each of the 427 locations. The data matrix $\mathbb{Y}$, reflecting precipitation, has 115 rows containing the annual summer-aggregated precipitation from 1902 to 2016 (lagged by one year from temperature) for each of the 427 locations. Analysis of precipitation versus summer temperatures lagged by 2 years, and temperatures from different seasons (Winter T; Summer P of the same

year) in the same year did not yield any significant bimodules after applying the FDR selection procedure.

5.5.2 Application of Search Procedure and Diagnostics

We ran BSP on the data with false discovery parameter $\alpha = 0.045$. (The selected α was the largest value in $\{0.01, 0.015, 0.02, \ldots 0.05\}$ having edge-error estimate under 0.1 based on 100 half-permuted datasets, see Section The edge-error estimates generally increases as α increases as shown in the following figure 5.1.

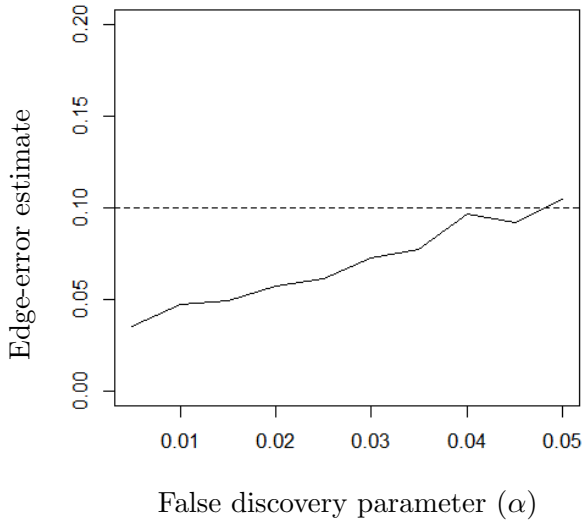


Figure 5.1: False discovery rates (FDR) for BSP results for the relationship between temperature (T) and precipitation(P) at significance levels ranging from 0.01 to 0.010. The largest value to be under the cutoff threshold at 0.10 is at 0.045

Our analysis does not directly take into consideration spatial data, but rather searches unsupervisedly for clusters of maximal correlations between related spatial processes, and then returns results that show the cross-variable spatial relationships across temperature and precipitation. Though temperature and precipitation are known to be spatially and temporally autocorrelated, and hence not completely indepedent, we assume that these effects are negligible compared to the true interactions between temperature and precipitation. Typical procedures in the analysis of spatially correlated processes make use of using correlograms and variograms to assess their depen-

dency (Legendre and Legendre, 2012). More detailed approaches to precisely parse out inter- and intra-correlations in T and P may be approached in future work.

BSP found five distinct bimodules, while the *effective number* of bimodules was three. After the filtering step, the two bimodules illustrated in Figure 5.2 and another bimodule with 80 temperature pixels and 5 precipitation pixels remained. We further omitted the latter bimodule since its precipitation pixels were same as those of bimodule B in Figure 5.2, but its temperature pixels were not geographically congruous.

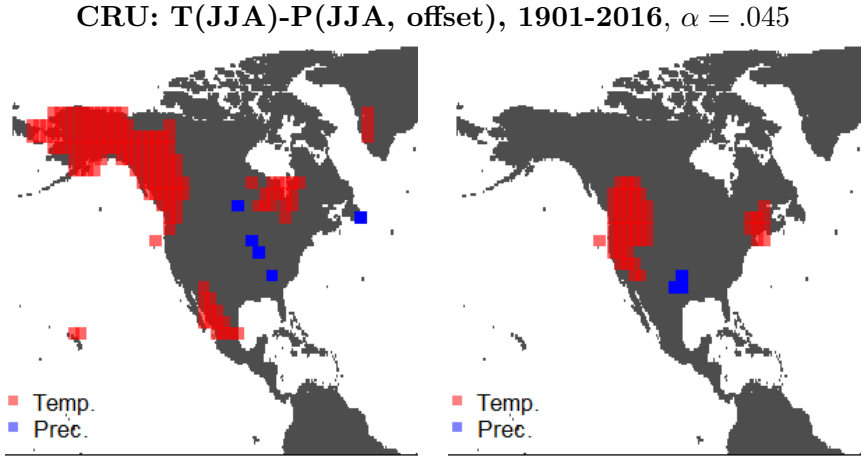


Figure 5.2: Bimodules of summer temperature and precipitation in North America from CRU observations from 1901-2016. The left bimodule (A) contains 149 temperature locations (pixels) and 6 precipitation locations. The right bimodule contains 53 temperature and 5 precipitation locations.

Temperature pixels in the two bimodules are situated distally from the precipitation pixels, but the temperature and precipitation pixels within a bimodule are form blocks of a contiguous geographical regions. Note that BSP did not use any location information while searching for these bimodules. However, the localization might also be because pixel nearby are correlated; though BSP typically accounts for these within-variable effects because of exchangeability in the data. We assume that these effects are not purely due to spatial dependence and instead from common sources of spatial origin.

The locations from the bimodules occupy large geographical areas on the map. The precipitation pixels from the bimodule on the left in Fig. 5.2 form a vertical stretch around the Eastern edge of the GP correlated with temperature pixels in large areas of land in the Pacific Northwest, Alaska, and Mexico. In the second bimodule Fig. 5.2 (right) precipitation pixels in the southern

Great Plains around Oklahoma is strongly correlated with temperature pixels in the Northwestern Great Plains. A an anomalously hot summer Oregon in one year in the Northwest suggests an anomalously rainy growing season in the following year in the Southern Great Plains. Pixel-wise positive correlations are confirmed in the following table 5.1

	P Pixel	Mean	SD
A	1	0.28	0.07
	2	0.27	0.06
	3	0.28	0.08
	4	0.27	0.08
	5	0.31	0.06
	6	0.30	0.08
	P Pixel	Mean	SD
B	1	0.31	0.04
	2	0.35	0.03
	3	0.29	0.04

Table 5.1: Average correlations per precipitation (P) pixel. for two bimodules A and B for climactic data (temperature and precipitation) in North America. Each entry yields a mean and standard deviation of the correlations each P pixel within the bimodule with every T pixel in the same bimodule. Results show all of the correlations are, at least on average, strongly positive.

The coastal proximity in all the temperature clusters suggest influences of oscillations in sea surface temperatures. Aforementioned patterns from both bimodules map to locations of agricultural productivity, such as in Oklahoma and Missouri (figure 5.2). The bimodules found by BSP only consider the magnitudes of correlations between the temperature and precipitation pixels. Upon further analysis of these bimodules we see that the significantly correlated temperature and precipitation pixels are positively correlated in the Great Plains region. These results agree with findings on concurrent T-P correlations in the Great Plains (Zhao and Khalil, 1993; Berg et al., 2015; Wang et al., 2019a) . Our findings demonstrate the utility of BSP in finding insights into remote correlations between precipitation and temperature in North America. Further research may build on these exploratory findings and create a model for the purpose of forecasting that can better predict growing-season precipitation in agriculturally productive regions around the world.

5.6 Application to Genomics (GTEX)

We describe results obtained from the application of bimodules to the problem of expression quantitative trait loci (eQTL) analysis. This application was primarily undertaken by Miheer

Dewaskar; more details can be found in the full article concerning this study (Dewaskar et al., 2020). The NIH funded GTEx Project has collected and created a large eQTL database containing genotype and expression data from postmortem tissues of human donors. A unique feature of this database is that it contains expression data from many tissues. We applied BSP and CONDOR to $p = 556304$ SNPs and $q = 26054$ thyroid expression measurements from $n = 574$ individuals.

We applied BSP to the thyroid eQTL data with false discovery parameter $\alpha = 0.03$ selected to keep the edge-error under 0.05 (Section 5.4.2). The search was initialized from singleton sets of all genes and half of the available SNPs, chosen at random. The effective number of bimodules was 3304 (using the Jaccard overlap method described in 3. The selected bimodules had SNP sets ranging in size from 1 to 1000 (median 1), and gene sets ranging in size from 1 to 100 (median 7).

5.6.1 Trans and Cis-eQTL Analysis

In order to assess potential biological utility of bimodules found by BSP, we compared the SNP-gene pairs in bimodules to those found by standard *cis*- and *trans*-eQTL analysis, studied the locations of the SNPs, and examined the gene sets for enrichment of known functional categories. Bimodules produced by CONDOR are similar to the SNP-gene pairs identified by *cis*- and *trans*-eQTL analysis. Table 5.2 compares these eQTL pairs with those found in bimodules identified by BSP. Cis-eQTL analysis considers only local SNP-gene pairs (improving detection power by reducing multiple testing), while trans-eQTL analysis and BSP do not use any information about locations of the SNPs and genes. Half of the pairs identified by cis-eQTL analysis and most of the pairs identified by trans-eQTL analysis appear in at least one bimodule.

Bimodules capture sub-networks of SNP-gene associations rather than individual eQTLs, and as such individual SNP-gene pairs in a bimodule need not be eQTLs. Table 5.2 shows that a significant fraction of BSP bimodules are not connected by either *cis*- or *trans*-eQTLs. The discovery of such bimodules suggests that the sub-networks identified by BSP cannot be found by standard eQTL analysis, and that these sub-networks can provide new insights and hypotheses for further study. To identify potentially new eQTLs using BSP, we examine bimodule connectivity under the combined set of cis- and trans-eQTLs. Around 300 local edges (i.e. the SNP is located within 1MB of the gene transcription start site) and 8.8K distal edges do not meet the correlation thresholds for *cis*- and *trans*-eQTL analysis, respectively, and should be investigated in future research.

distance type	% eQTLs found among bimodules
trans analysis	84%
cis analysis	51%

Table 5.2: Comparison of BSP and standard eQTL analysis. A gene-SNP pair is said to be found among a collection bimodules if the gene and SNP are both part of some common bimodule.

5.6.2 Genomic locations and Ontology

We studied the chromosomal location and proximity of SNPs and genes from bimodules found by BSP and CONDOR. While CONDOR uses genomic locations as part of the cis-eQTL analysis, BSP does not make use of location information. Genetic control of expression is often enriched in a region local to the gene (Consortium et al., 2017). All CONDOR clusters and almost all bimodules, have at least one local SNP-gene pair, wherein the SNP is located within 1MB of the gene transcription start site. All SNPs and all but two (Chr. 8 and 9) of the genes from CONDOR clusters were located on Chromosome 6. The SNPs and genes from the bimodules were distributed across all 23 chromosomes:

The Gene Ontology (GO) database contains a curated collection of gene sets that are known to be associated with different biological functions (Consortium, 2014; Botstein et al., 2000; Rhee et al., 2008). The topGO (Alexa and Rahnenfuhrer, 2018) package determines placement of each sets in enriched GO sets. For each of the 145 "large" gene sets(>8) bimodules, we used topGO to assess the biological processes of B. We retained results with significant BH q-values ($\alpha = .05$). Of the 145 gene sets considered, 18 had significant overlap with one or more biological process. We further refine the enrichment results using text analysis in the `quanteda` package (Benoit et al., 2018). The most significant GO terms for BSP are be found in the following table. In bimodule (indexed) 1, for example, many of the ontologies associated with the discovery appear to relate antigen processing to the discovered genes; in bimodule 3, detection of chemical stimulus; in bimodule 11, regulation of cellular processes; ion response in bimodule 14. The repeated occurrences of these enrichments speak to the power of discovering biologically relevant results using BSP.

Bimodule	Ontologies
1	antigen(9), immune(9), processing(8), presentation (8), response(8)
6	regulation(11), viral(10), entry(7), response (6)
9	regulation(10) , biosynthetic(8) , process(6)
11	regulation(10) , biosynthetic(8) , process(6)
14	ion(10) , response(9) , cellular(9)

Table 5.3: Text analysis of of the gene ontology results for resulting bimodules. The ontology keywords with greater than 5 occurences were filtered. The analysis was conduected using the R package `quanteda`.

CHAPTER 6
Future Work

In this chapter, we discuss potential avenues of future work. The two primary directions in future work are the multivariate normal weighted stochastic blockmodel and a null model for spatial networks. We also discuss potential technical extensions to the intertemporal community detection method described in chapter 3. These future directions tie together several common themes form the previous sections. Notably the presence of a global background set of unclustered nodes and network representations of spatial relationships.

There are two broad parts of this chapter that describe ongoing as well as future work. In the first part (Section 6.1) we highlight the more immediate "ongoing" work that represents the next step of the applications of `SBANM` in Chapter 2. In this first section, I outline a *naive* method for prediction based on Mahalanobis distances in order to assess prediction error. In the following sections, I describe future directions in self-looping networks outlined in Chapter 3 in Section 6.2 and also intertemporal community detection described in Chapter 4 in Section 6.3. In Section 6.4, we describe a detailed extension of the null model for self-looping networks described in Chapter 2 for spatial networks.

6.1 Ongoing Extensions to `SBANM`

The development of `SBANM` opens up a bevy of methodological avenues. One immediate next step is to expand the study of PNC data to neuroimaging and genomics data. Such work is currently in progress for the PNC study to identify potentially jointly model neural and genetic influences in addition to symptoms. Another direction is in assessing significance or predictive power of the imputed clusters. More generally, these in-group and out-of-group interactions are related to mixed effects models for multimodal weighted networks that may serve as another perspective in the study longitudinal analysis of networks (Snijders, 2005).

We propose a simple extension of the SBANM method to assess prediction errors. This extension has been implemented on data, but its methdological and theoretical justifications should be further explored in future work.

6.1.1 Cross Validation

We describe an algorithm of cross validation (CV) to assess prediction error. This algorithm itself may be used for other community detection methods, particularly variational inference-based ones that may naturally allow for Mahalanobis distances to be factored into the clustering mechanism.

1. For n samples split into $n/2$ sets for training and test set
2. Apply SBANM on the *training set* $\mathbf{X}^{train}$ and obtain $\mathbf{\Theta}_{train} = \left(\boldsymbol{\mu}_q, \mathbf{\Sigma}_q\right)_{q:q\leq Q}$ for fixed Q blocks (one being noise), obtain the memberships s $\mathbf{Z}^{train}$
3. For every observation $\mathbf{X}_i^{test}$ in the *test set*:
   (a) For every member j, find the distance between group q between the edge (i, j)
   (b) Find the group q that has the closest Mahalanobis Distance for every member-edge j, tabulate all (i, j) across n^{test} groups with the smallest distances.
   (c) Initialize membership vector $\widehat{\tau}_i$ with the proportion of tabulated minimum -distance memberships across *all other edges*
   (d) Apply one round of the E-Step of SBANM to determine the memberships $\mathbf{Z}^{test}$, given the already-estimated parameters $\mathbf{\Theta}_{train}$.
4. Compare the memberships $\mathbf{Z}^{test}$ to the estimated $\mathbf{Z}^{train}$ and assess CV prediction error for memberships
5. Repeat the above steps but reverse the roles of training and test set

6.1.2 Cross Validation Results

We assess the cross-validation error for the *early adult* segment of the PNC data. the sample consists of 1863 subjects. Earlier tests found that the optimal selection for Q was 4 for this sample.

For simplicity , we only focus on the noise block NB as it is the one that is most easily identifiable across different runs, which are specifically labeled in the model estimation algorithm. We use a Jaccard score to assess the match between the cross validation hold-out sets and the *true value*. The ground truth in this case does not exist, but we refer to the *true value* as the imputed memberships from a run of the algorithm in the *training set*. Note that the training set and test set are applied twice, and flipped. In future work we will conduct a simulation study to evaluate this method with the presence of ground truth.

The Jaccard score is defined as the ratio of the cardinality total intersecting members from two sets A and B, which is similarly defined in Chapter 4. a higher score signifies higher agreement. We use the Jaccard score between the *predicted* and *true* (as assessed form the *other symmetric experiment*). To contrast against the symmetric Jaccard index, we also use the one-sided predictive error which is the ratio of the overlap $O^{test} = A^{pred,test} \bigcap A^{true,test}$ to the true test set $A^{true,test}$, the prediction rate only measures if the true members are recovered, and does not take into account the false positives.

$$\text{Pred}\% = \frac{|O^{test}|}{|A^{true,test}|}.$$

The symmetric Jaccard score Jaccard($A^{true,test}, A^{pred,test}$), as defined in Chapter 3, on the other hand, accounts for false positives as well as true recoveries:

Trial	Train_{True}	Test_{True}	Test_{Pred}	Overlap	Jaccard	Pred%
1a	30	29	39	24	.55	.83
1b	29	30	56	28	.48	.93
2a	2	5	9	4	.40	.80
2b	5	2	20	0	0	0
3a	4	4	12	3	.23	.75
3b	4	4	13	4	.31	1
4a	3	8	14	1	.50	.12
4b	8	3	43	1	.20	.33
5a	6	7	30	6	.19	.86
5b	7	6	16	1	.50	.17

Table 6.1: Estimates and ground truths of each half-sample CV split for `SBANM` applied to PNC early adults

The prediction rates do seem to indicate that most of the true members in NB are found, however, there seems to be a lot of false positives.

6.1.3 Group Sizes of Each Run

Here I provide the most basic summary statistics of each half-run. The sizes of each NB are all very small compared to the total sample size (1863). Moreover, a major "control" group of around 700-800 S_1 is also consistently found. CV errors for these groups should also be assessed in following work. In general, however, these results appear to be promising in the consistent discovery of similar clusters across evenly-split training/test sets for the `SBANM` algorithm.

Trial 1				
	ρ_q(a)	| Gps(a) |	ρ_q(b)	| Gps(b) |
NB	0	30	0	29
S_1	49	751	51	754
S_2	62	148	70	147
S_3	100	2	81	2

Trial 2				
	ρ_q(a)	| Gps(a) |	ρ_q(b)	| Gps(b) |
NB	0	2	0	5
S_1	40	69	11	83
S_2	52	775	49	731
S_3	64	85	50	113

Trial 3				
	ρ_q(a)	| Gps(a) |	ρ_q(b)	| Gps(b) |
NB	0	4	0	4
S_1	50	762	50	749
S_2	41	67	19	73
S_3	57	98	57	106

Trial 4				
	ρ_q(a)	| Gps(a) |	ρ_q(b)	| Gps(b) |
NB	0	3	0	8
S_1	46	757	47	699
S_2	25	52	55	217
S_3	53	119	83	8

Trial 5				
	ρ_q(a)	| Gps(a) |	ρ_q(b)	| Gps(b) |
NB	0	6	0	7
S_1	51	752	51	764
S_2	52	45	18	76
S_3	60	128	57	85

Table 6.2: Clustering characteristics of *training-sets* for the 5 trials shown above.a) and (b) respectively represent the flipped training sets (which serves as the test set in a subsequent analysis) that comprise half of the total sample. ρ_q for (a) and (a) are the estimated correlations (times 100). Each ρ_q for every block designated NB are set to zero. |Gps(a)| and |Gps(b)| denote the estimated block sizes for each block.

6.2 Limitations and Further Research in Self-Looping Networks

Several limitations arise from the proposed CCME-SL method and application described in Chapter 3. Though mostly unsupervised, there are several tuning parameters that must be specified, such as the overlap parameter and the α threshold for the p-value. Sensitivity analysis (see Fig. 3 in Supporting Information) over a range of p-values and overlap thresholds demonstrates that results do not change much when tuning parameters are tweaked. The post-processing step for determining the nodal communities does not capture the extent of the monocentricity that arises

from areas that we expect to exhibit these properties, such as Multnomah County (Portland) and Hennepin County (Minneapolis).

Two major issues that we plan to explore in future work are:

1. **Resolution limit:** In future work we plan to explore the notion of resolution limit in the context of CCME and its judgement of significant communities. In brief, consider the simpler setting of unweighted networks. In an unweighted network the null model corresponding to an empirically observed network is created by placing an edge between two vertices u and v with a probability proportional to $d_u d_v / d_T$. For two sets A and B with degrees $d(A)$ and $d(B)$, $d(A) \times d(B) \ll d_T$ will cause the null model to find any presence of an edge between these sets to be "surprising" and cause A and B to be merged into the same community. This is a well-known issue with algorithms underpinned by a null model, as is the case for modularity detection. Similar issues must also arise with CCME-like algorithms. One method to address this would be to incorporate a resolution parameter γ as in the context of the unweighted case, which we now briefly describe (and refer the reader to work in (Reichardt and Bornholdt, 2006) or (Fortunato, 2010, Section 6.C) for more details). In the unweighted case, the natural null model is the configuration model which preserves degrees. This model implies that the null random graph model gives the probability of the existence of an edge between two vertices u and v as in (3.3). To accommodate for and deal with the issue of the resolution limit, the reference model can be modified so that

$$\mathbb{P}(\tilde{A}_{uv} = 1) = \gamma \frac{d_u d_v}{d_T}$$

   As one increases $\gamma \uparrow \infty$, this implies that we expect non-trivial connectivity between nodes and thus connectivity within subsets to pass "higher bars" before being judged significant. We are exploring similar ideas in the context of the weighted case, including connections between modifications of the reference null model and the corresponding Markov stability of diffusion processes on the null models (Lambiotte et al., 2008).

2. **Spatial null models:** In current work we are developing null models which also take into account the spatial component i.e. null models that directly include the spatial component

and preserve various functionals such as the degree and the strength. We hope that this new approach will give more accurate results, mitigating issues like that of CCME-SL finding connections between neighboring counties surprising purely because of the resolution limit, since CCME-SL does not directly incorporate the notion that in spatial systems "neighboring counties have a higher propensity to connect".

Several fruitful directions of further research emerge from this study of commuting regions. This study documents the influence of self loops in spatial networks describing complex patterns arising from the collective behavior of many individuals. As network data in these realms become increasingly available, we expect the emergence of more methods accounting for self-looping behavior in networks. We plan to extend the methodology proposed in this study to directed networks so as to model the orientation of commuter flows. A further extension of the method described in this study is to use a temporal model to measure change in communities across time. Another avenue is to investigate the characteristics of power-law distribution of populations that are embedded within the commuting networks and how distributional characteristics of power laws interplay in community detection.

This research extends such a methodology to a more general setting of spatial networks that characterizes collective behaviors. Such networks often represent agglomerations of particles that are influenced by both core and peripheral elements. The method of community detection in networks with self-loops may be applied to a variety of spatially-constrained human mobility networks which naturally exhibit significant self-looping characteristics, such as human migration behaviors.

This method may find applications in other domains as well. In neuroimaging, one such application may be in analyzing the epicenter-spreading proliferation of biomarkers such as *tau*, which is significantly linked to Alzheimer's Disease. Research has revealed that tau develops along a trajectory of concentric spatial accumulation that aggregates at a seed region. Mapping such behavior in brain networks may find a suitable implemention in community detection on strongly self-looping graphs.

6.3 Future Work in Intertemporal Community Extraction

We aim to expand on the work in intertemporal community detection as outlined in Chapter 5. An immediatre next step would be to design and implement a simulation study to assess model performance, and to deternine how to select the optiimal energy barrier U. One other further direction is in more rigorously accounting for the false discovery rate control and accounting for the temporal dependence catering to a variety of situations, such as clusters with varying rates of change in connectivity across time. Presently, the time-series clustering method accounts for repeated counts of significance as well as postulated directions or connectivity trajectory. Future explorations may (1) account for vector autoregressive models or (2) test for changepoints for differing trends in connectivity.

6.4 Spatial Null Model

In future work, we propose a spatial null model to describe Hubs and Cliques in networks that representing spatial relationships. This model may be viewed as an extension of some hypotheses regarding network structure in the discussion of the CCME-SL in Chapter 3. Present literature describe 'latent position models' using probability models in the following form ((Hoff et al., 2002b).

$$\mathbb{P}(A_{uv} = 1) = \sigma \exp\left(\frac{-\text{dist}(u, v)}{\rho}\right) \tag{6.1}$$

For some parameters ϕ, τ and latent variables z_i, z_j. Some prior work has been focused on latent position network models (Hoff et al., 2002a). Define z_{uv} as the distance between two nodes u and v. Although z_{uv} can actually be any sort of latent graph underlying the primary graph.

We propose a rough outline of a spatial null model on a single weighted graph $\mathbf{X}$ whose weights are described as X_{uv} between nodes u, v. Moreover, there is an underlying adjacency matrix A whose entries are 0 or 1 and describe the presence of an edge A_{uv} between two nodes. In addition, between two nodes u, v we use $\text{dist}(u, v)$ for encoding their geographical distance. The global resolution parameter may be calculated in another grid search or by heuristic assumption i.e. $\tau = 1$ or $\tau = 1/s_T$. One way of determining τ is through a grid search that minimizies the sum mean squared error for each estimate.

We posit two different null models that reflect two different "viewpoints" for an observer that is located at some spatial point that is represented by a node. These null models may be used in a variety of ways: we propose that they can be used to detect different types of communities arising from networks that represent spatially relational structures.

We expand on the postulation in chapter 3 that spatial networks observe inherently two different types of clustering formations: hub-spoke structures, wherein most nodes in a given cluster connect strongly to a central node, and clique-like communities wehre all nodes are connected to each other. We prospectively propose a method based on variational inference for the detection and classification of these different types of communities.

6.4.1 Local Null Model

The assumption of the local null model is that the network driven by same-location point processes (self loops), that radiate outwards. Observer at u knows:

1. distance $\text{dist}(u, v)$ between u and v
2. value of weight X_{uu} at point u, because the weight at the immediate point can be observed
3. total strength S_u of node u, implying, since that, since the observer also knows every edge A_{uv} and degree $\deg(u)$, they also know the average weight
4. sample variance κ of the weights X_{uv}; intuitively overall "neighborhood" variation (surrounding u) is known to the viewer, though they do not know the individual components.

Because the position of the observer is localized, they *do not know* the (1) binary edge A_{uv} for all v that is connected to u, (2) exact weight X_{uv} for $v \neq u$, (3) frequency $f(X_{uv})$ of weight X_{uv}.

In the local model, one has a 'general idea' of the weight structure across the entire spatial domain by presupposing knowlege of S_u. This assumption is analogously used in the configuration model: total weights across each node are fixed parameters. To use an analogy as a motivating example : suppose a node represents a traveler who has just moved into a new city represented by a network of neighborhoods. They generally know what and where a neighborhood is like, and they know where *their* neighborhood is , but they are not totally sure where their neighborhood connects to another neighborhood, even though know all bordering neighborhoods. Another analogy for the

above heuristics is: one is standing in a subway, they know what train schedule is in the given station, and they know the scale (represented by S_u) as well as pace, or complexity, of the station within the context of the system (represented by κ_u). However, where trains are coming and going in *other* stations, the observer does not know.

We treat the conditional expectation $\mathbb{E}_u[X_{uv}|u,v]$ as a random variable with some arbitary distribution specific to u with (1) mean μ_u and (2) variance σ_u^2. The best approximate guess for each weight X_{uv} is its conditional expectation given the distance, thus the estimate for all weights X_{uv} from node u is Y_u multiplied by a spatial decay term that decreases with distance. Such a term, at its simplest, can be represented by an exponential covariance function. The expected value of a weight given point u is, for local scaling parameters ρ_u unique to each node and for global resolution parameter τ:

$$\mathbb{E}_u[X_{uv}|u] = \frac{\mu_v}{\tau} \exp\left(- \frac{\text{dist}(u,v)}{\rho_u} \right) \tag{6.2}$$

The variance of weight X_{uv} connecting two points u, v is similarly constructed and is evocative of the Matern covariance used often in spatial statistics, implying stationarity. Now we calculate the parameters and probabilities using the constraints. Prior work has used the weighted configuration model in spatial null models (Ruzzenenti et al., 2012) , but have not taken into account any constraints based on variances using existing work on Gaussian process covariance functions.

1. **Preserves conditional expected strength**

   As in configuration model, preserve the strengths, or sum of weights. For $S_u^* = S_u - W_{uu}$ as the strength excluding self loop:

   $$\mathbb{E}_u[\widehat{S_u^*}|u] = S_u - W_{uu}$$

   for weights centered around a point u, we assume that all points have the same expectation scaled by distance *except* for the point u itself, which is known as the "observer" is standing directly above that point.

   $$\mathbb{E}[\widehat{S}_u^*|u] \propto \mu_u \cdot g(\text{Area}(u), \rho_u)$$

where $g(\cdot)$ is some arbitrary continuous function that is composed of the area surrounding point u and the global resolution parameter ρ.

2. **Preserves sum of conditional variances**

   Define κ as the *known* sample variance of the system, multiplied by n. The value is known, intuitively, (despite the sum of its parts being treated as "unknown") because the observer has a general idea of the size and complexity of the spatial system (i.e. the total activity of the train station), but not the details of its individual parts. We set the constraint

$$\sum_{v:v\sim u} \text{Var}(X_{uv}|u) \approx \sum_{v:v\sim u} (X_{uv} - \mathbb{E}[X_{uv}|u])^2 := \kappa_u$$

   Such a condition presumes that the spatial spread of variances is proportional to the sample variance, given the knowledge of X_{uv} from being centered at u.

Since the observer is presumed to know the strength as well as the degrees of u, as well as the self-loop, so one choice for the estimate of μ_u given u, $\mathbb{E}_u[X_{uv}|u]$ is $\widehat{\mu}_u = S_u - X_{uu}/(\deg(u) - 1)$. Therefore, the fitted value for each weight between u, v, if the observer is located at u, is

$$\widehat{X}_{uv} = \begin{cases} \frac{\widehat{\mu}_u}{\tau} \exp\left(- \frac{\text{dist}(u,v)}{\rho_u} \right) & \text{if } u \neq v \\ X_{uu} & \text{if } u = v \end{cases}.$$

6.4.2 Global Model

Assumptions for the global model are identical to those of the local spatial model but the key difference is that the probability between edges u and v must be estimated using global assumptions. The observer at u does not have any local knowledge of u but instead knows the global estimates for probabilities and expected edges. Like in the local model: the observer takes as given: distance $\text{dist}(u, v)$ between u and v, total strength S_u and degree $\deg(u)$, of node u sample variance κ of the weights X_{uv}. However, they **do not know** the (1) **edge connection** A_{uv}, (2) **value of weight** X_{uu} at point u, (3) exact weight X_{uv} for $v \neq u$, (4) frequency $f(X_{uv})$ of weight X_{uv}.

The primary difference between the local and global model is that in the global model, the estimated probability of edge connection is instead a function of τ^{global} and ρ^{global}. The fitted value

for each edge X_{uv} is then the expected value under the global null distribution, observed from the vantage point of u: $\mathbb{E}_u[X_{uv}] = \mathbb{E}_u[X_{uv}|A_{uv}] \cdot \mathbb{P}(A_{uv})$.

6.4.3 Testing for Hubness and Cliqueness

I outline a plan to use variational EM to alternate between estimation of the model parameters for both the local and global parameters in future work. The descriptions here are exploratory and outline potential steps that we may take. The variational parameters are the sets of nodes $B_{(k)}$ (at a given iteration k), obtained by iterative testing in a similar way as outlined in the previous chapters. The model parameters are the global and local parameters μ_{uv}, ρ_{uv} for each node u, v and entire-network parameters τ for both the global and local null models. We estimate the probability of a candidate set B (at iteration k) being a *Hub* (H), or a *Clique* C. We represent this probability of node u being a hub as variational parameter P_u, which converges to zero or one. For a given set B, we construct a test statistic $D_u(B)$ that represents the difference between hubness and cliqueness:

$$D_u(B) = H(u, B) - C(u, B)$$

for $B = \{\text{Candidate set, inclusive of } u\}$, and $B^- = \{\text{Candidate set, not inclusive of } u\}$

$$H(u, B) = \sum_{v:v\in B} X_{uv}$$
$$C(u, B) = \sum_{w:w\in B^-} \sum_{v:v\in B^-} X_{wv}$$

But for "spoke" u, it is either in a hub with probability $1 - P_u$ or a part of a polycentric clique with probability P_u, so the test statistic is

$$S(u, B) = P_u^D(B)C(u, B) + (1 - P_u^D(B))H(u, B)$$

In the E-step, the memberships are estimated by performing iterative testing on the statistic $S(u, B)$. At each iterative step k, a candidate set $B_{(k)}$ is performed with its edge between all nodes

in the network:

$$p(u, B) = \mathbb{P}(S(u, B) > S(u, B, \mathcal{G})).$$

where $\mathcal{G}$ represents the random graph under the null model. The p-value is calculated as some approximation of the difference between the cliqueness $C(u, B)$ and hubness $H(u, B)$, whose statistics are respectively estimated from the global and local models:

$$\mathbb{E}S(u, B) = P_u^D \mathbb{E}C(u, B) + (1 - P_u^D)\mathbb{E}H(u, B).$$

Each conditional expectation term inside the summation can be calculated, with respect to global and local assumptions. This is iterated, alternating with the M-step wherein the local and global parameters $\widehat{\mu}_{uv}^{\text{global}}$ and $\widehat{\mu}_u$ (as well as associated variances) are estimated, until membership variables and hubness probabilities P_u become stable. To calculate the variational variable P_u which represents the probability that set B is a hub but not a clique.

$$P_u^D(B) = \mathbb{P}(D_u(B) > \mathbb{E}D_u(B)).$$

These calculations are then performed for each seed set in a similar way as the iterative testing schemes in the previous chapters. These outlined steps will be fleshed out in future work.

BIBLIOGRAPHY

Abbe, E. (2017). Community detection and stochastic block models: recent developments.

Adler, R. F., Gu, G., Wang, J.-J., Huffman, G. J., Curtis, S., and Bolvin, D. (2008). Relationships between global precipitation and surface temperature on interannual and longer timescales (1979–2006). *Journal of Geophysical Research: Atmospheres*, 113(D22).

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2007). Mixed membership stochastic blockmodels.

Alexa, A. and Rahnenfuhrer, J. (2018). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.34.0.

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Allman, E. S., Matias, C., and Rhodes, J. A. (2011). Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference*, 141(5):1719–1736.

Ambroise, C. and Matias, C. (2010). New consistent and asymptotically normal estimators for random graph mixture models.

Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E., and Vogelstein, J. T. (2020). Inference for multiple heterogeneous networks with a common invariant subspace.

Arroyo Relión, J. D., Kessler, D., Levina, E., and Taylor, S. F. (2019). Network classification with applications to brain connectomics. *The Annals of Applied Statistics*, 13(3).

Assem, H., Xu, L., Buda, T. S., and O'Sullivan, D. (2016). Spatio-Temporal Clustering Approach for Detecting Functional Regions in Cities. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 370–377.

Austwick, M. Z., O'Brien, O., Strano, E., and Viana, M. (2013). The Structure of Spatial Networks and Communities in Bicycle Sharing Systems. *PLOS ONE*, 8(9):e74685.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102.

Barthélemy, M. (2014). *Spatial networks*. Springer.

Batty, M. (2013). *The new science of cities*. MIT Press.

Beckett, S. J. (2016). Improved community detection in weighted bipartite networks. *Royal Society open science*, 3(1):140536.

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Berg, A., Lintner, B. R., Findell, K., Seneviratne, S. I., van den Hurk, B., Ducharne, A., Chéruy, F., Hagemann, S., Lawrence, D. M., Malyshev, S., Meier, A., and Gentine, P. (2015). Interannual coupling between summertime surface temperature and precipitation over land: Processes and implications for climate change. *Journal of Climate*, 28(3):1308–1328.

Bickel, P. and Chen., A. (2009). A nonparametric view of network models and newman– girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blevins, A. S., Kim, J. Z., and Bassett, D. S. (2021). Variability in higher order structure of noise added to weighted networks.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308.

Bodwin, K., Zhang, K., and Nobel, A. (2015). A testing-based approach to the discovery of differentially correlated variable sets.

Bollobás, B. (1980). A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316. bibtex[publisher=Academic Press].

Botstein, D., Cherry, J. M., Ashburner, M., Ball, C., Blake, J., Butler, H., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000). Gene ontology: tool for the unification of biology. *Nat genet*, 25(1):25–9.

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

Calkins, M. E., Merikangas, K. R., Moore, T. M., Burstein, M., Behr, M. A., Satterthwaite, T. D., Ruparel, K., Wolf, D. H., Roalf, D. R., Mentch, F. D., Qiu, H., Chiavacci, R., Connolly, J. J., Sleiman, P. M., Gur, R. C., Hakonarson, H., and Gur, R. E. (2015). The philadelphia neurodevelopmental cohort: constructing a deep phenotyping collaborative. *Journal of Child Psychology and Psychiatry*, 56(12):1356–1369.

Calkins, M. E., Moore, T. M., Merikangas, K. R., Burstein, M., Satterthwaite, T. D., Bilker, W. B., Ruparel, K., Chiavacci, R., Wolf, D. H., Mentch, F., Qiu, H., Connolly, J. J., Sleiman, P. A., Hakonarson, H., Gur, R. C., and Gur, R. E. (2014). The psychosis spectrum in a young u.s. community sample: findings from the philadelphia neurodevelopmental cohort. *World Psychiatry*, 13(3):296–305.

Calkins, M. E., Moore, T. M., Satterthwaite, T. D., Wolf, D. H., Turetsky, B. I., Roalf, D. R., Merikangas, K. R., Ruparel, K., Kohler, C. G., Gur, R. C., and Gur, R. E. (2017). Persistence of psychosis spectrum symptoms in the philadelphia neurodevelopmental cohort: a prospective two-year follow-up. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, 16(1):62—76.

Cannon, T. D., Yu, C., Addington, J., Bearden, C. E., Cadenhead, K. S., Cornblatt, B. A., Heinssen, R., Jeffries, C. D., Mathalon, D. H., McGlashan, T. H., Perkins, D. O., Seidman, L. J., Tsuang, M. T., Walker, E. F., Woods, S. W., and Kattan, M. W. (2016). An individualized risk calculator for research in prodromal psychosis. *American Journal of Psychiatry*, 173(10):980–988. PMID: 27363508.

Carlen, J., de Dios Pont, J., Mentus, C., Chang, S.-S., Wang, S., and Porter, M. A. (2019). Role detection in bicycle-sharing networks using multilayer stochastic block models.

Cazabet, R., Borgnat, P., and Jensen, P. (2017a). Using Degree Constrained Gravity Null-Models to understand the structure of journeys' networks in Bicycle Sharing Systems. In *ESANN 2017 - European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Cazabet, R., Borgnat, P., and Jensen, P. (2017b). Using degree constrained gravity null-models to understand the structure of journeys' networks in Bicycle Sharing Systems. In *ESANN 2017 - European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium.

Chakraborty, T., Dalmia, A., Mukherjee, A., and Ganguly, N. (2017). Metrics for community analysis: A survey. *ACM Comput. Surv.*, 50(4):54:1–54:37.

Cheng, W., Shi, Y., Zhang, X., and Wang, W. (2015). Fast and robust group-wise eqtl mapping using sparse graphical models. *BMC bioinformatics*, 16(1):2.

Cho, Y.-S., Steeg, G. V., and Galstyan, A. (2011). Co-evolution of selection and influence in social networks.

Citibike (2019). Citibike. https://www.citibikenyc.com. Accessed: 2019-12-24.

Clark, L., Watson, D., and Reynolds, S. (1995). Diagnosis and classification of psychopathology: challenges to the current system and future directions. *Annual review of psychology*, 46:121—153.

Clauset, A., E J Newman, M., and Moore, C. (2005). Finding community structure in very large networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 70:066111.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Consortium, G. et al. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204.

Consortium, G. O. (2014). Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056.

Costa, A. and Hansen, P. (2014). A locally optimal hierarchical divisive heuristic for bipartite modularity maximization. *Optimization Letters*, 8(3):903–917.

Cupo, L., McIlwaine, S. V., Daneault, J.-G., Malla, A. K., Iyer, S. N., Joober, R., and Shah, J. L. (2021). Timing, Distribution, and Relationship Between Nonpsychotic and Subthreshold Psychotic Symptoms Prior to Emergence of a First Episode of Psychosis. *Schizophrenia Bulletin*. sbaa183.

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

Dewaskar, M., Palowitch, J., He, M., Love, M. I., and Nobel, A. (2020). Finding stable groups of cross-correlated features in multi-view data.

Divvy (2019). Divvy data.

Divvy (2019). Divvy gbfs data. `https://gbfs.divvybikes.com/gbfs/en/station_status.json`. Accessed: 2019-10-28.

Dixon, P. M. and Pechmann, J. H. K. (2008). A statistical test to show negligible trend: Reply. *Ecology*, 89(5):1473–1473.

Du, Z., Yang, B., and Liu, J. (2017). Understanding the spatial and temporal activity patterns of subway mobility flows. *arXiv preprint arXiv:1702.02456*.

Dunn, O. J. (1959). Estimation of the medians for dependent variables. *Annals of Mathematical Statistics*, 30(1):192–197.

Durrett, R. (2007). *Random graph dynamics*, volume 200. Cambridge university press Cambridge.

Ertöz, L., Steinbach, M., and Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *SDM*.

Ewens, W. J. (2003). On estimating p values by the monte carlo method. *American Journal of Human Genetics*, 72(2):496–498.

Faghih-Imani, A. (2014). Analysing bicycle sharing system user destination choice preferences : An investigation of chicago's divvy system.

Faghih-Imani, A. and Eluru, N. (2015). Analysing bicycle-sharing system user destination choice preferences: Chicago's divvy system. *Journal of Transport Geography*, 44:53 – 64.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3-5):75–174.

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

Fosdick, B. K., Larremore, D. B., Nishimura, J., and Ugander, J. (2018). Configuring random graph models with fixed degree sequences. *SIAM Review*, 60:315–355.

Fowler, S., C., Jensen, L., and Rhubart, D. (2018). Assessing U.S. labor market delineations for containment, economic core, and wage correlation.

Freund, D., Henderson, S. G., and Shmoys, D. B. (2018). Minimizing multimodular functions and allocating capacity in bike-sharing systems.

Fujishima, S., Fujiwara, N., Akiyama, Y., Shibasaki, R., and Sakuramachi, R. (2019). The size distribution of 'cities' delineated with a network theory-based method and mobile phone gps data.

Gast, N., Massonnet, G., Reijsbergen, D., and Tribastone, M. (2015). Probabilistic forecasts of bike-sharing systems for journey planning. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 703–712, New York, NY, USA. ACM.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Good, B. H., de Montjoye, Y.-A., and Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81:046106.

Greenfield, J. (2018). After a 2016 slump, divvy turned a record profit in 2017. *Chicago Reader*.

Hagler, Y. (2009). Defining US megaregions. Technical report, Regional Planning Association, New York.

Han, Y. and Goetz, S. J. (2019). Overlapping labour market areas based on link communities. *Papers in Regional Science*, 98(1):539–553.

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 (Statistics in Society)*, 170(2):301–354.

Hao, Z., Hao, F., Singh, V. P., and Zhang, X. (2018). Quantifying the relationship between compound dry and hot events and el niño–southern oscillation (enso) at the global scale. *Journal of Hydrology*, 567:332 – 338.

Harris, I., Jones, P., Osborn, T., and Lister, D. (2014). Updated high-resolution grids of monthly climatic observations – the cru ts3.10 dataset. *International Journal of Climatology*, 34(3):623–642.

He, M., Glasser, J., Bhamidi, S., and Kaza, N. (2020a). Intertemporal community detection in human mobility networks.

He, M., Glasser, J., Pritchard, N., Bhamidi, S., and Kaza, N. (2020b). Demarcating geographic regions using community detection in commuting networks with significant self-loops. *PLOS ONE*, 15(4):e0230941.

He, M., Lu, D., Xu, J., and Xavier, R. M. (2021). Community detection in weighted multilayer networks with ambient noise.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002a). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002b). Latent space approaches to social network analysis.

Hoffman, M., Blei, D. M., Wang, C., and Paisley, J. (2012). Stochastic variational inference.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137.

Holme, P. (2015). Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9).

Huang, L., Yang, Y., Gao, H., Zhao, X., and Du, Z. (2018). Comparing community detection algorithms in transport networks via points of interest. *IEEE Access*, 6:29729–29738.

Huang, Y., Wuchty, S., Ferdig, M. T., and Przytycka, T. M. (2009). Graph theoretical approach to study eqtl: a case study of plasmodium falciparum. *Bioinformatics*, 25(12):i15–i20.

Isserman, A. (2005). In the national interest: Defining rural and urban correctly in research and public policy. *International Regional Science Review*, 28(4):465–499.

Jaakkola, T. S. (2000). Tutorial on variational approximation methods. In *IN ADVANCED MEAN FIELD METHODS: THEORY AND PRACTICE*, pages 129–159. MIT Press.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241–272.

Jones, M. and Paasi, A. (2013). Regional World(s): Advancing the Geography of Regions. *Regional Studies*, 47(1):1–5.

Kahn, R., Sommer, I., Murray, R., Meyer-Lindenberg, A., Weinberger, D., Cannon, T., O'Donovan, M., Correll, C., Kane, J., Van Os, J., and Insel, T. (2015). Schizophrenia. *Nature Reviews Disease Primers*, 1.

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

Kendell, R. and Jablensky, A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American Journal of Psychiatry*, 160(1):4–12. PMID: 12505793.

Kim, K., Chun, Y., and Kim, H. (2017). p-Functional Clusters Location Problem for Detecting Spatial Clusters with Covering Approach. *Geographical Analysis*, 49(1):101–121.

Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477.

Lambiotte, R., Delvenne, J.-C., and Barahona, M. (2008). Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*.

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PLOS ONE*, 6:1–18.

Latouche, P., Birmelé, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336.

Legendre, P. and Legendre, L. (2012). Chapter 1 - complex ecological data sets. In Legendre, P. and Legendre, L., editors, *Numerical Ecology*, volume 24 of *Developments in Environmental Modelling*, pages 1 – 57. Elsevier.

Levin, K., Athreya, A., Tang, M., Lyzinski, V., Park, Y., and Priebe, C. E. (2019). A central limit theorem for an omnibus embedding of multiple random graphs and implications for multiscale network inference.

Lewis, J. B., Poole, K., Rosenthal, H., Boche, A., Rudkin, A., and Sonnet, L. (2020). Voteview: Congressional roll-call votes database.

Liu, J., Sun, L., Chen, W., and Xiong, H. (2016). Rebalancing bike sharing systems: A multi-source data smart optimization. In *KDD '16*.

Liu, S., Wang, S., and Krishnan, R. (2014). Persistent community detection in dynamic social networks. In Tseng, V. S., Ho, T. B., Zhou, Z.-H., Chen, A. L. P., and Kao, H.-Y., editors, *Advances in Knowledge Discovery and Data Mining*, pages 78–89, Cham. Springer International Publishing.

Liu, W., Suzumura, T., Ji, H., and Hu, G. (2018). Finding overlapping communities in multilayer networks. *PLOS ONE*, 13(4):1–22.

Liu, X., Gong, L., Gong, Y., and Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography*, 43:78–90.

Liu, X. and Murata, T. (2010). An efficient algorithm for optimizing bipartite modularity in bipartite networks. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 14(4):408–415.

Livneh, B. and Hoerling, M. P. (2016). The physics of drought in the u.s. central great plains. *Journal of Climate*, 29(18):6783–6804.

Madden, R. A. and Williams, J. (1978). The correlation between temperature and precipitation in the united states and europe. *Monthly Weather Review*, 106(1):142–147.

Mariadassou, M., Robin, S., and Vacher, C. (2010). Uncovering latent structure in valued graphs: A variational approach. *Ann. Appl. Stat.*, 4(2):715–742.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913.

Mathews, H., Mayya, V., Volfovsky, A., and Reeves, G. (2019). Gaussian mixture models for stochastic block models with non-vanishing noise.

Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141.

Mayya, V. and Reeves, G. (2019). Mutual information in community detection with covariate information and correlated networks.

McCabe, S. D., Lin, D.-Y., and Love, M. I. (2019). Consistency and overfitting of multi-omics methods on experimental data. *Brief Bioinform*.

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4):628–641.

Menichetti, G., Remondini, D., Panzarasa, P., Mondragón, R. J., and Bianconi, G. (2014). Weighted multiplex networks. *PLOS ONE*, 9(6):1–8.

Mercado, P., Tudisco, F., and Hein, M. (2019). Spectral clustering of signed graphs via matrix power means.

Nelson, G. D. and Rae, A. (2016). An Economic Geography of the United States: From Commutes to Megaregions. *PLOS ONE*, 11(11):e0166083.

Newman, M. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351.

Newman, M. (2018a). *Networks*. Oxford university press.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E. and Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564–9569.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3). arXiv: physics/0605087.

Newman, M. E. J. (2018b). Estimating network structure from unreliable measurements. *Physical Review E*, 98(6).

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

NYC Taxi and Limousine Commission (2020). Tlc trip record data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2020-1-20.

Nystuen, J. D. and Dacey, M. F. (1961). A graph theory interpretation of nodal regions. *Papers of the Regional Science Association*, 7(1):29–42.

Open BUS (2019). The open bus. `https://www.theopenbus.com/raw-data.html`. Accessed: 2019-12-23.

Opsahl, T. and Panzarasa, P. (2009). Clustering in weighted networks. *Social Networks*, 31(2):155–163.

Paasi, A. (2013). Regional Planning and the Mobilization of 'Regional Identity': From Bounded Spaces to Relational Complexity. *Regional Studies*, 47(8):1206–1219.

Palowitch, J., Bhamidi, S., and Nobel, A. B. (2018). The Continuous Configuration Model: A Null for Community Detection on Weighted Networks. *Journal of Machine Learning Research*, 18:1–48.

Pan, C., Luo, J., Zhang, J., and Li, X. (2019). BiModule: biclique modularity strategy for identifying transcription factor and microRNA co-regulatory modules. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Pappalardo, L., Barlacchi, G., Pellungrini, R., and Simini, F. (2019). Human mobility from theory to practice:data, models and applications. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 1311–1312, New York, NY, USA. ACM.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, 8(1).

Patel, P. V., Gianoulis, T. A., Bjornson, R. D., Yip, K. Y., Engelman, D. M., and Gerstein, M. B. (2010). Analysis of membrane proteins in metagenomics: Networks of correlated environmental features and protein families. *Genome Research*, 20(7):960–971.

Patuelli, R., Reggiani, A., Nijkamp, P., and Bade, F.-J. (2010). The evolution of the commuting network in germany: Spatial and connectivity patterns. *Journal of Transport and Land Use*, 2(3):5–37.

Paul, S. and Chen, Y. (2015). Community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *preprint*.

Paul, S. and Chen, Y. (2018). A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. *preprint*.

Peel, L., Larremore, D. B., and Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*, 3(5).

Peixoto, T. P. (2013). Parsimonious module inference in large networks. *Physical review letters*, 110(14):148701.

Peixoto, T. P. (2014). Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047.

Peixoto, T. P. (2017). Bayesian stochastic blockmodeling.

Peixoto, T. P. (2018). Nonparametric weighted stochastic block models. *Phys. Rev. E*, 97:012306.

Pendem, P. (2019). Maximizing ridership in bicycle-sharing systems using empirical data and stochastic models. `public.kenan-flagler.unc.edu/MSOM2017_3_0302`. Accessed: 2019-10-20.

Pesantez-Cabrera, P. and Kalyanaraman, A. (2016). Detecting communities in biological bipartite networks. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 98–107.

Pike, A., editor (2013). *'Whither regional studies?'*. Routledge, New York, NY.

Plane, D. A. (1981). The geography of urban commuting fields: some empirical evidence from New England. *The Professional Geographer*, 33(2):182–188.

Platig, J., Castaldi, P. J., DeMeo, D., and Quackenbush, J. (2016). Bipartite community structure of eqtls. *PLoS Computational Biology*, 12(9):e1005033.

Pucher, B. M., Zeleznik, O. A., and Thallinger, G. G. (2019). Comparison and evaluation of integrative methods for the analysis of multilevel omics data: a study based on simulated and experimental cancer data. *Briefings in bioinformatics*, 20(2):671–681.

Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3120–3128, USA. Curran Associates Inc.

Ranganath, R., Tran, D., and Blei, D. (2016). Hierarchical variational models. *Proceedings of the 33 rd International Conference on Machine Learning,*, 18.

Reades, J., Calabrese, F., and Ratti, C. (2009). Eigenplaces: Analysing cities using the space–time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836.

Reeves, G., Mayya, V., and Volfovsky, A. (2019). The geometry of community detection via the mmse matrix.

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

Ren, Y., Ercsey-Ravasz, M., Wang, P. P., González, M. C., and Toroczkai, Z. (2014). Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nature communications*, 5:5347.

Rhee, S. Y., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509–515.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39(4):1878–1915.

Ruzzenenti, F., Picciolo, F., Basosi, R., and Garlaschelli, D. (2012). Spatial effects in real networks: Measures, null models, and applications. *Physical Review E*, 86(6).

Salter-Townshend, M. and Murphy, T. B. (2013). Variational bayesian inference for the latent position cluster model for network data. *Computational Statistics and Data Analysis*, 57(1):661–671.

Sarzynska, M., Leicht, E. A., Chowell, G., and Porter, M. A. (2015). Null models for community detection in spatially embedded, temporal networks. *Journal of Complex Networks*, 4(3):363–406.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680.

Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392):96.

Snijders, T. A. B. (2005). Models for longitudinal network data. In *Models and Methods in Social Network Analysis*, pages 215–247. University Press.

Stanley, N., Shai, S., Taylor, D., and Mucha, P. J. (2015). Clustering network layers with the strata multilayer stochastic block model. *CoRR*, abs/1507.01826.

Tini, G., Marchetti, L., Priami, C., and Scott-Boyer, M.-P. (2019). Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in bioinformatics*, 20(4):1269–1279.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240.

Tong, D. and Plane, D. A. (2014). A New Spatial Optimization Perspective on the Delineation of Metropolitan and Micropolitan Statistical Areas. *Geographical Analysis*, 46(3):230–249.

Tong, Y., Chen, Y., Zhou, Z., Chen, L., Wang, J., Yang, Q., Ye, J., and Lv, W. (2017). The simpler the better: A unified approach to predicting original taxi demands based on large-scale online platforms. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1653–1662, New York, NY, USA. Association for Computing Machinery.

Tropp, J. A. (2006). Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051.

van der Hofstad, R. (2016). *Random Graphs and Complex Networks: Volume 1*. Cambridge University Press, Cambridge, 1 edition edition.

van Praag, H. M. (2000). Nosologomania: a disorder of psychiatry. *The World Journal of Biological Psychiatry*, 1(3):151–158.

Wang, B., Luo, X., Yang, Y.-M., Sun, W., Cane, M. A., Cai, W., Yeh, S.-W., and Liu, J. (2019a). Historical change of el niño properties sheds light on future changes of extreme el niño. *Proceedings of the National Academy of Sciences*, 116(45):22512–22517.

Wang, S., Arroyo, J., Vogelstein, J. T., and Priebe, C. E. (2019b). Joint embedding of graphs.

Wheeler, S. (2013). Regions, Megaregions, and Sustainability.

Wilson, J. D., Palowitch, J., Bhamidi, S., and Nobel, A. B. (2016). Community extraction in multilayer networks with heterogeneous community structure.

Wilson, J. D., Stevens, N. T., and Woodall, W. H. (2019). Modeling and detecting change in temporal networks via the degree corrected stochastic block model. *Quality and Reliability Engineering International*, 35(5):1363–1378.

Wilson, J. D., Wang, S., Mucha, P. J., Bhamidi, S., and Nobel, A. B. (2014). A testing based extraction algorithm for identifying significant communities in networks. *Annals of Applied Statistics*, 8(1):1853–1891.

Wu, X., Liu, Q., and Jiang, R. (2009). Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics*, 25(1):98–104. Publisher: Oxford Academic.

Xiang, J., Tang, Y.-N., Gao, Y.-Y., Zhang, Y., Deng, K., Xu, X.-K., and Hu, K. (2015). Multi-resolution community detection based on generalized self-loop rescaling strategy. *Physica A: Statistical Mechanics and its Applications*, 432:127 – 139.

Xie, X.-F. and Wang, Z. J. (2018). Examining travel patterns and characteristics in a bikesharing network and implications for data-driven decision supports: Case study in the washington dc area. *Journal of Transport Geography*, 71:84 – 102.

Yan, X., Shalizi, C., Jensen, J. E., Krzakala, F., Moore, C., Zdeborova, L., Zhang, P., and Zhu., Y. (2014). Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment*, 5:05–07.

Young, J.-G., Cantwell, G. T., and Newman, M. E. J. (2020). Bayesian inference of network structure from unreliable data. *Journal of Complex Networks*, 8(6).

Zhan, X., Qian, X., and Ukkusuri, S. V. (2016). A graph-based approach to measuring the efficiency of an urban taxi service system. *IEEE Transactions on Intelligent Transportation Systems*, 17(9):2479–2489.

Zhao, W. and Khalil, M. A. K. (1993). The relationship between precipitation and temperature over the contiguous united states. *Journal of Climate*, 6(6):1232–1236.

Zhao, Y., Levina, E., and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.*, 40(4):2266–2292.

Zhong, C., Arisona, S. M., Huang, X., Batty, M., and Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. 28:2178–2199.

Zhou, X. (2015). Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in chicago. *PLOS ONE*, 10(10):1–20.

Zhou, Y.-H., Barry, W. T., and Wright, F. A. (2013). Empirical pathway analysis, without permutation. *Biostatistics*, 14(3):573–585.