Statistical Reasoning in Network Data

by

Youjin Lee

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

January, 2019

Abstract

Networks are collections of nodes, which can represent entities like people, genes, or brain regions, and ties between pairs of nodes, which represent various forms of connection, e.g. social relationships, between them. The study of networks is booming in biology, economics, statistics, psychology, physics, computer science, social science, public health, and beyond. Despite the increased interest in network data and its application, methods do not yet exist to answer many types of statistical and causal questions about observations collected from networks.

In this dissertation, we illustrate an unacknowledged problem for statistical methods using network data, namely *network dependence*, and propose a test for the existence of such dependence. We demonstrate how this kind of dependence affects the validity of statistical inference. In particular, one of the most important sources of data on cardiovascular disease epidemiology, the Framingham Heart Study, is shown to exhibit dependence that could lead to false statistical conclusions. We also propose a network dependence test that

overcomes the high-dimensional structure of network data.

Many researchers interested in social networks in public health and social science are ultimately interested in causal inference on certain collective behaviors or health outcomes observed over the whole network – such as the causal effect of a certain vaccination plan on the overall rate of infections, or the causal effect of an online viral marketing program on the sales of products. In the last part of the dissertation, we focus on one of those questions that aims to identify the most influential subjects in networks.

ABSTRACT

Primary Readers:

Elizabeth L. Ogburn (Primary Advisor)
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Carl Latkin
Professor
Department of Health, Behavior, and Society
Johns Hopkins Bloomberg School of Public Health

Ilya Shpitser
Assistant Professor
Department of Computer Science
Johns Hopkins Whiting School of Engineering

Abhirup Datta
Assistant Professor
Department of Biotstatistics
Johns Hopkins Bloomberg School of Public Health

Alternative Readers:

Elizabeth Stuart
Professor
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health

Michael A. Rosenblum
Associate Professor
Department of Biotstatistics
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

I cannot imagine how this work would be written without my advisor, Betsy Ogburn. This first word in the acknowledgement reminds me of very first moment I knocked the door of her office. From then she led me to the world of social network and causal inference from my total ignorance to the topic and no research experience. I always loved to talk and work with her throughout my PhD program. Her insightful comments on research and writing have guided me to move in the right direction, but yet she always left some room for improvements with independent and creative thinking.

I am glad to say a big thank you to my undergraduate advisor, Myung-Hee Cho Paik at Seoul National University, South Korea. Her support and encouragement brought me here to this wonderful environment of Johns Hopkins Biostatistics. My thanks also go out to the support from Kwanjeong Educational Foundation.

I am very thankful to thesis readers, Carl Latkin from the Department of Health, Behavior, and Society, Abhirup Datta from the Department of Bio-

statistics, and Ilya Shpitser from the Department of Computer Science. Special thanks to Ilya Shpitser for his philosophical guidance toward causal inference. I am also grateful to Elizabeth Stuart and Michael Rosenblum for their consideration and time toward my thesis.

I would like to thank causal inference working group and Survival, Longitudinal And Multivariate data (SLAM) working group at the Department of Biostatistics. Working groups within the department always kept me motivated to learn and discuss interesting research topics. Along with weekly departmental seminar, these study groups gave me an opportunity to connect to many researchers from different institutions, which would definitely enrich my career in the future.

I would like to thank Mei-Cheng Wang for her valuable advice and support. She led me to view the data with statistical perspectives, and it helped me to think about research problems from the data. Especially, she invited me to the research about delivery and reproductive history for women, which diversified my research. I really enjoyed collaboration with Rajeshwari Sundaram from National Institutes of Health and Li Liu from the Department of Population, Family, and Reproductive Health, Johns Hopkins Bloomberg School of Public Health.

My experience at NeuroData lab in the Department of Biomedical Engineering mentored by Joshua Vogelstein has opened my eyes to other part of network

science. His devotion and passion toward research always motivated me. My research with him would not be possible without great mentor, Cencheng Shen now at University of Delaware.

I also thank my friends (too many to list them all!) for spending fun and memorable time with me in Baltimore. I cannot imagine my life in Baltimore without them. I would like to thank my colleagues in the Department of Biostatistics. Discussion about our research and our lives nurtured my everyday life. A very special thank you to Mary Joy, a departmental academic administrator. Without her, I could have not registered for the class, arranged my oral examination, and presented my doctoral defense. I appreciate her help for all of those.

I am deeply thankful to my family and my four grandparents for their unconditional love and support. They always respect me and support my life. I always think how lucky I am to have their love. I have saved my last word of this acknowledgement for my dear husband Cory Cho, who has been with me all these years and just started a new chapter of our lives. With these grateful moments in my mind, I am ready to start our new chapter.

January 2019

Youjin Lee

Contents

List of Tables

List of Figures

LIST OF FIGURES

Chapter 1

Introduction

1.1 Statistical problems in network data

In many scientific and public health studies, observations are collected from subjects who are related to each other as members of one or a small number of social networks. For example, subjects are often sampled from one or small number of schools, hospitals, geographic areas, or online communities, where they may be connected via *social tie*s or *edge*s such as being friends or sharing the same teacher or medical provider. These subjects, often called *nodes* of the network, are interacting with each other while their features or behaviors are changing over time, dependent on others' through social ties.

In public health, social network data has received a lot of attention largely due to the interest in the ways social interactions or collective behaviors among

humans affect health outcomes in populations (Kaufman, 2017). There has been much research on the relationship between social networks and mortality (Berkman and Syme, 1979), mental health (Kawachi and Berkman, 2001; Russell and Cutrona, 1991), infectious diseases (Eubank et al., 2004; Christley et al., 2005), and behavioral changes (Voorhees et al., 2005; Centola, 2011). For the last decade, a series of influential papers by Christakis and Fowler purport to demonstrate that health outcomes, behaviors and attitudes, like obesity (Christakis and Fowler, 2007), smoking (Christakis and Fowler, 2008) or happiness (Fowler and Christakis, 2008), spread through social networks. Implicitly or explicitly these relationships are causal (Berkman and Syme, 1979; Kawachi and Berkman, 2001; Russell and Cutrona, 1991).

Despite increased interest in network data in public health and social science, however, we found a lack of valid and approachable statistical methods for observations collected from network nodes, and standard statistical methods developed for independent observations have been widely used for network data. Causal inference with observations from network nodes is especially challenging due to the requirement for high-dimensional data. To illustrate, to infer a causal statement, e.g. "my friend's weight gain *causes* my weight gain", using observational data from a single network requires observing longitudinal data of all the relevant observations, e.g. my and my friend's weights over time and all the confounding factors affecting these two outcomes, which explain all

the existing causal relationships involved. In this setting, the number of observations required explodes over time, and in most cases it is impossible to collect the kind of real-time data required. Even if we had access to the requisite data, the resulting model will be high-dimensional and often too big to fit in practice.

Often core research questions raised in social network studies require causal concepts. We introduce one of them in the dissertation: "who is the most influential subject in a social network?". To answer this question, most researchers defined influence only through descriptive features of the underlying network or presumed diffusion model, even though some of these researchers inherently attempted to identify *causally* influential subjects, who would exert a substantial causal effect on the whole network.

This dissertation does not provide a perfect solution to overcome all of the aforementioned challenges; instead we demonstrate the necessity for thorough diagnostics on statistical inference for network data and also for rigorous causal understanding of social dynamics.

1.2 Organizational overview

In this dissertation, we present statistical methods for network data in three parts. The first part, presented in Chapter 2 and Chapter 3, introduces the concept of network dependence and proposes a method to test for such de-

pendence. We further demonstrate that network dependence can lead to invalid and biased statistical inference. In addition to simulation studies, we apply our test for network dependence to several published papers that use the Framingham Heart Study (FHS) data.

In the second part of the dissertation, presented in Chapter 4, we propose a new approach to test for network dependence in the presence of high-dimensional nodal attributes. To overcome model-based approaches and structural obstacles in network data, we use distance-based correlations applied to the network embeddings, which yield a theoretically consistent test statistic under mild graph distributional assumptions. Through simulations, we demonstrate that the test works well for many popular network models. We apply our distance-based tests on the neuronal network and implement independence test between synapse connectivity and each neuron's position.

While the first two parts of the dissertation are mostly about testing for dependence in network data and the impact of such dependency on general statistical inference, the last part illustrates how causal inference on network nodes' outcomes can answer a question raised in the study of networks across many disciplines. In Chapter 5, we define the influence of each node in a network through its causal impact on the collective outcomes across the network. Chapter 5 uses a specific statistical model, detailed in Appendix B, but suggests other approaches beyond specific model-based inference.

CHAPTER 1. INTRODUCTION

We present proofs and additional simulations for testing network dependence under high-dimensional setting in Appendix A. In Appendix B we discuss the details of causal inference on collective outcomes using causal graphical model called chain graph.

Chapter 2

Testing Network and Spatial Autocorrelation

Testing for dependence has been a well-established component of spatial statistical analyses for decades. In particular, several popular test statistics have desirable properties for testing for the presence of spatial autocorrelation in continuous variables. In this chapter we propose two contributions to the literature on tests for autocorrelation. First, we propose a new test for autocorrelation in categorical variables. While some methods currently exist for assessing spatial autocorrelation in categorical variables, the most popular method is unwieldy, somewhat ad hoc, and fails to provide grounds for a single omnibus test. Second, we discuss the importance of testing for autocorrelation in network, rather than spatial, data, motivated by applications in social net-

work data. We demonstrate that existing tests for autocorrelation in spatial data for continuous variables and our new test for categorical variables can both be used in the network setting.

This is a joint work in collaboration with Elizabeth Ogburn.

2.1 Introduction

In studies using spatial data, researchers routinely test for spatial dependence before proceeding with statistical analysis (Legendre, 1993; Lichstein et al., 2002; Diniz-Filho et al., 2003; F Dormann et al., 2007). Spatial dependence is usually assumed to have an autocorrelation structure, whereby pairwise correlations between data points are a function of the geographic distance between the two observations (Cliff and Ord, 1968, 1972). Because autocorrelation is a violation of the assumption of *independent and identically distributed* (i.i.d.) observations or residuals required by most standard statistical models and hypothesis tests (Legendre, 1993; Anselin et al., 1996; Lennon, 2000), testing for spatial autocorrelation is a necessary step for valid statistical inference using spatial data. For continuous random variables, the most popular tests are based on Moran's I statistic (Moran, 1948) and Geary's C statistic (Geary, 1954). For categorical random variables, however, available tests based on join count analysis (Cliff and Ord, 1970) are unwieldy and fail to provide a single

omnibus test of dependence.

Taking temporal dependence into account is similarly widely practiced in time series settings. But other kinds of statistical dependence are routinely ignored. In many public health and social science studies, observations are collected from individuals who are members of one or a small number of social networks within the target population, often for reasons of convenience or expense. For example, individuals may be sampled from one or a small number of schools, institutions, or online communities, where they may be connected by ties such as being related to one another; being friends, neighbors, acquaintances, or coworkers; or sharing the same teacher or medical provider. If individuals in a sample are related to one another in these ways, they may not furnish independent observations, which undermines the assumption of i.i.d. data on which most statistical analyses in the literature rely.

In the literature on spatial and temporal dependence, dependence is often implicitly assumed to be the result of latent traits that are more similar for observations that are close than for distant observations. This *latent variable dependence* (Ogburn, 2017) is likely to be present in many network contexts as well. In networks, ties often present opportunities to transmit traits or information from one node to another, and such direct transmission will result in *dependence due to direct transmission* (Ogburn, 2017) that is informed by the underlying network structure. In general, both of these sources of dependence

result in positive pairwise correlations that tend to be larger for pairs of observations from nodes that are close in the network and smaller for observations from nodes that are distant in the network. Network distance is usually measured by geodesic distance, which is a count of the number of edges along the shortest path between two nodes. This is analogous to spatial and temporal dependence, which are generally thought to be inversely related to (Euclidean) distance.

Despite increasing interest in and availability of social network data, there is a dearth of valid statistical methods to account network dependence. Although many statistical methods exist for dealing with dependent data, almost all of these methods are intended for spatial or temporal data or, more broadly, for observations with positions in $\mathbb{R}^k$ and dependence that is related to Euclidean distance between pairs of points. The topology of a network is very different from that of Euclidean space, and many of the methods that have been developed to accommodate Euclidean dependence are not appropriate for network dependence. The most important difference is the distribution of pairwise distances which, in Euclidean settings, is usually assumed to skew towards larger distances as the sample grows, with the maximum distance tending to infinity with n. In social networks, on the other hand, pairwise distances tend to be concentrated on shorter distances and may be bounded from above. However, as we elaborate in Section 2.2, methods that have been used to test for

spatial dependence can be adapted and applied to network data.

A few papers have proposed using Moran's I in network settings: to confirm suspected dependence in network (Black, 1992; Long et al., 2015), to identify appropriate weight matrices for regression models (Butts et al., 2008), or to find the largest correlation for dimension reduction (Fouss et al., 2016). Many variables of interest in social network studies are categorical, for example group affiliations (Kossinets and Watts, 2006), personality (Adamic et al., 2003), or ethnicity (Lewis et al., 2008). Join count analysis has been recently used for testing autocorrelation in categorical outcomes observed from social networks (e.g. Long et al. (2015)). Farber et al. (2009) proposed a more elegant test for categorical network data and explored its performance in data generated from a linear spatial autoregression (SAR) model. As far as we are aware, all of the previous work assumes that the network data were generated from SAR models, and none of this previous work has considered the performance of autocorrelation tests for more general network settings.

In this chapter we propose a new test statistic that generalizes Moran's I for categorical random variables. We also propose to use both Moran's I and our new test for categorical data to assess the hypothesis of independence among observations sampled from a single social network (or a small number of networks). We assume that any dependence is monotonically inversely related to the pairwise distance between nodes, but otherwise we make no assumptions

on the structure of the dependence. These tests allow researchers to assess the validity of i.i.d statistical methods, and are therefore the first step towards correcting the practice of defaulting to i.i.d. methods even when data may exhibit network dependence.

2.2 Methods

2.2.1 Moran's I

Moran's I takes as input an n-vector of continuous random variables and an $n \times n$ weighted distance matrix $\mathbf{W}$, where entry w_{ij} is a non-negative, non-increasing function of the Euclidean distance between observations i and j. Moran's I is expected to be large when pairs of observations with greater w values (i.e. closer in space) have larger correlations than observations with smaller w values (i.e. farther in space). The choice of non-increasing function used to construct $\mathbf{W}$ is informed by background knowledge about how dependence decays with distance; it affects the power but not the validity of tests of independence based on Moran's I. The asymptotic distribution of Moran's I under independence is well established (Sen, 1976) and can be used to construct hypothesis tests of the null hypothesis of independence. Geary's c (Geary, 1954) is another statistic commonly used to test for spatial autocorrelation (Fortin

et al., 1989; Lam et al., 2002; da Silva et al., 2008); it is very similar to Moran's I but more sensitive to local, rather than global, dependence. We focus on Moran's I in what follows because our interest is in global, rather than local, dependence. Because of the similarities between the two statistics, Geary's c can be adapted to network settings much as we adapt Moran's I.

Let Y be a continuous variable of interest and y_i be its realized observation for each of n units ($i = 1, 2, \ldots, n$). Each observation is associated with a location, traditionally in space but we will extend this to networks. Let W be a weight matrix signifying closeness between the units, e.g. a matrix of pairwise Euclidean distances for spatial data or an adjacency matrix for network data. (The entries A_{ij} in the adjacency matrix A for a network are indicators of whether nodes i and j share a tie.) Then Moran's I is defined as follows:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\big(y_i - \bar{y}\big)\big(y_j - \bar{y}\big)}{S_0 \sum_{i=1}^{n}\big(y_i - \bar{y}\big)^2/n}, \tag{2.1}$$

where $S_0 = \sum_{i=1}^{n}(w_{ij} + w_{ji})/2$ and $\bar{y} = \sum_{i=1}^{n} y_i/n$. Under independence, the pairwise products $(y_i - \bar{y})(y_j - \bar{y})$ are each expected to be close to zero. On the other hand, under network dependence adjacent pairs are more likely to have similar values than non-adjacent pairs, and $(y_i - \bar{y})(y_j - \bar{y})$ will tend to be relatively large for the upweighted adjacent pairs; therefore, Moran's I is expected to be larger in the presence of network dependence than under the null hypothesis

of independence.

The exact mean μ_I and variance σ_I^2 of Moran's I under independence are given in Sen (1976) and Getis and Ord (1992). The standardized statistic $I_{std} := (I\mu_I)/\sqrt{\sigma_I^2}$ is asymptotically normally distributed under mild conditions on W and Y (Sen, 1976). Using the known asymptotic distribution of the test statistic under the null permits hypothesis tests of independence using the normal approximation. For network data we propose a permutation test based on permuting the Y values associated with each node while holding the network topology constant. Setting $w_{ij} = 0$ for all non-adjacent pairs of nodes results in increased variability of I relative to spatial data, and therefore the normal approximation may require larger sample sizes to be valid for network data compared to spatial data. The permutation test is valid regardless of the distribution of W and Y and for small sample sizes.

2.2.2 New methods for categorical random variables

For a K-level categorical random variable, join count statistics compare the number of adjacent pairs falling into the same category to the expected number of such pairs under independence, essentially performing K separate hypothesis tests. As the number of categories increases, join count analyses become

quite cumbersome. Furthermore, they only consider adjacent observations, thereby throwing away potentially informative pairs of observations that are non-adjacent but may still exhibit dependence. Finally, the K separate hypothesis tests required for a join count analysis are non-independent and it is not entirely clear how to correct for multiple testing. To overcome this last limitation, Farber et al. (2015) proposed a single test statistic that combines the K separate joint count statistics.

Instead of extending join count analysis, we propose a new statistic for categorical observations using the logic of Moran's I. This has two advantages over the proposal of Farber et al. (2015): it incorporates information from discordant, in addition to concordant, pairs and it weights kinds of pairs according to their probability under the null. To illustrate, under network dependence adjacent nodes are more likely to have concordant outcomes and less likely to have discordant outcomes than they would be under independence. We operationalize independence as random distribution of the outcome across the network, holding fixed the marginal probabilities of each category. The less likely a concordant pair (under independence), the more evidence it provides for network dependence, and the less likely a discordant pair (under independence), the more evidence it provides against network dependence. Using this rationale, a test statistic should put higher weight on more unlikely observations. The following is our proposed test statistic:

$$\Phi = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\big\{2\mathbf{I}(y_i = y_j) - 1\big\}/p_{y_i}p_{y_j}}{S_0}, \tag{2.2}$$

where $p_{y_i} = P(Y = y_i)$, $p_{y_j} = P(Y = y_j)$, and $S_0 = \sum_{i=1}^{n}(w_{ij} + w_{ji})/2$. The term $(2\mathbf{I}(y_i = y_j) - 1) \in \{-1, 1\}$ allows concordant pairs to provide evidence for dependence and discordant pairs to provide evidence against dependence. The product of the proportions p_{y_i} and p_{y_j} in the denominator ensures that more unlikely pairs contribute more to the statistic. As the true population proportion is generally unknown, $\{p_k : k = 1, ..., K\}$ should be estimated by sample proportions for each category.

The first and second moment of Φ are derived in the Appendix 2.6.1. Asymptotic normality of the statistic Φ under the null can also be proven based on the asymptotic behavior of statistics defined as weighted sums under some constraints. For more details see Appendix 2.6.2. For binary observations, which can be viewed as categorical or continuous, our proposed statistic has the desirable property that the standardized version of Φ is equivalent to the standardized Moran's I.

2.2.3 Choosing the weight matrix W

Tests for spatial dependence take Euclidean distances (usually in $\mathbb{R}^2$ or $\mathbb{R}^3$) as inputs into the weight matrix $\mathbf{W}$. In networks, the entries in $\mathbf{W}$ can be comprised of any non-increasing function of geodesic distance for the purposes of the tests for network autocorrelation that we describe below, but for robustness we use the adjacency matrix $\mathbf{A}$ for $\mathbf{W}$, where A_{ij} is an indicator of nodes i and j sharing a tie. The choice of $\mathbf{W} = \mathbf{A}$ puts weight 1 on pairs of observations at a distance of 1 and weight 0 otherwise. In many spatial settings, subject matter expertise can facilitate informed choices of weights for $\mathbf{W}$ (e.g. Smouse and Peakall 1999; Overmars et al. 2003), but it is harder to imagine settings where researchers have information about how dependence decays with geodesic network distance. In particular, dependence due to direct transmission is transitive: dependence between two nodes at a distance of 2 is through their mutual contact. This kind of dependence would be related to the number, and not just length, of paths between two nodes. It may also be possible to construct distance metrics that incorporate information about the number and length of paths between two nodes, but this is beyond the scope of this chapter. In general in the presence of network dependence adjacent nodes have the greatest expected correlations; therefore $\mathbf{W} = \mathbf{A}$ is a valid choice in all settings. Of course, if we have knowledge of the true dependence mechanism, using a weight matrix that incorporate this information will increase power.

2.3 Simulations

In Section 2.3.1, we demonstrate the validity and performance of our new statistic, Φ, for testing spatial autocorrelation in categorical variables. In Section 2.3.2, we demonstrate the performance of Moran's I and Φ for testing for network dependence.

2.3.1 Testing for spatial autocorrelation in categorical variables

We replicated one of the data generating settings used by Farber et al. (2015) and implemented permuation-based tests of spatial dependence using Φ. First, we generated a binary weight matrix $\mathbf{W}$ with entries w_{ij} indicating whether regions i and j are adjacent. The number of neighbors (d_i) for each site i was randomly generated through $d_i = 1 + \text{Binomial}(2(d-1), 0.5)$. We simulated 500 independent replicates of $n = 100$ observations under each of four different settings, with $d = 3, 5, 7, 10$.

We then used $\mathbf{W}$ to generate a continuous, autocorrelated variable:

$$Y^* = (I_n - \rho\mathbf{W})^{-1}\epsilon$$

where I_n is a $n \times n$ identity matrix, and $\epsilon_i \sim N(0, 1)$ and ρ controls the amount

of dependence. We applied cutoffs based on the $(0.25, 0.5, 0.75)$ quantiles of each simulated dataset to convert Y^* into categorical observations $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ having $K = 4$ categories.

Figure 2.1 presents the simulation results. It shows that under the null ($\rho = 0$), the rejection rate is close to the nominal level of $\alpha = 0.05$ and that power to detect dependence increases with ρ.

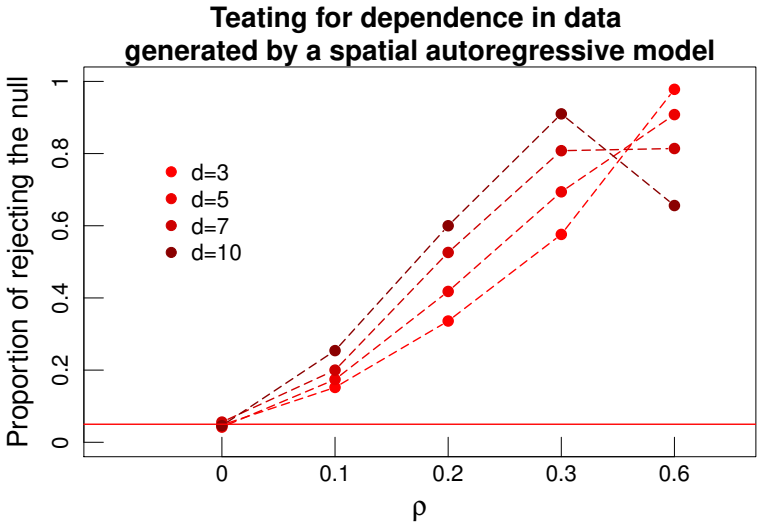


Figure 2.1: Permutation tests based on Φ. Dependence increases as ρ increases, and the y-axis is the proportion of 500 independent simulations in which the test rejected the null hypothesis of independence.

We also simulated data under a spatially correlated error model (F Dormann et al., 2007), using a continuous weight matrix estimated from real spatial data. We used the longitude and latitude of 473 U.S. power generating facilities (Papadogeorgou, 2017; Papadogeorgou et al., 2016) to construct a Euclidean distance matrix $\mathbf{D} = [d_{ij}]$, where d_{ij} is the Euclidean distance between

facilities i and j, based on which we constructed a weight matrix $\mathbf{\Pi} = [\pi_{ij}]$ where $\pi_{ij} = \exp(-qd_{ij}/\max(\{d_{ij} : i, j = 1, 2, \ldots, n\}))$. The amount of dependence is controlled by q. For each of four settings (no dependence, $q = 100, 50, 25$) we simulated $n = 473$ observations $\mathbf{Y}^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)$ 500 times according to the following model:

$$\mathbf{Y}^* \sim \mathbf{B}^T \xi, \tag{2.3}$$

where $\mathbf{\Pi} = \mathbf{B}^T\mathbf{B}$ and $\xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Finally, we applied cutoffs based on the $(0.1, 0.3, 0.6, 0.85)$ quantiles of each simulated dataset to convert Y^* into categorical observations $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ having $K = 5$ categories.

We calculated Φ two different ways: using the correct weight matrix, $\mathbf{\Pi}$, and using an estimated weight matrix $\mathbf{W}$:

$$W_{ij} = \max(D/d_{ij}, 10)$$

$$W_{ii} = 0,$$

where $D = \max_{i,j} d_{ij}$ ensures that the smallest weight is 1. The resulting weights w_{ij} are inversely proportional to the Euclidean distance between facility i and j, but truncated at 10. The percentage of $w_{ij} = 10$, i.e., the percentage of truncated weights, is about 12%.

Figure 2.2 shows that tests of independence based on Φ using $\mathbf{W}$ have increasing power as dependence increases, while tests using the true weight ma-

trix Π have nearly perfect power under all three alternatives.

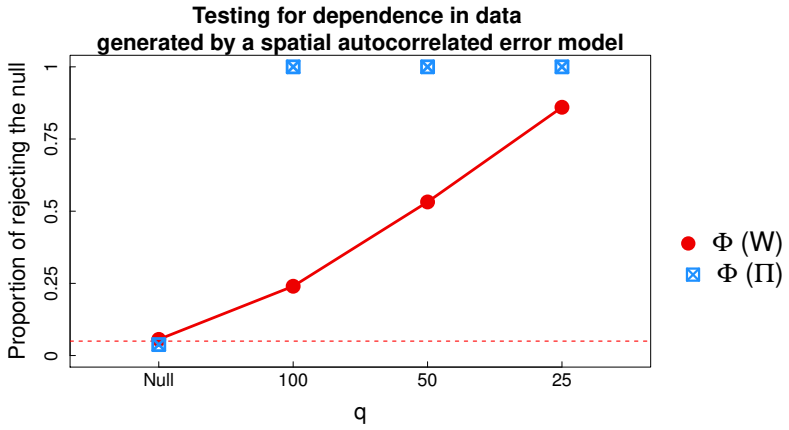


Figure 2.2: Permutation tests based on Φ. Dependence increases as q increases, and the y-axis is the proportion of 500 independent simulations in which the test rejected the null hypothesis of independence.

2.3.2 Testing for network dependence

In this section we simulate continuous and categorical random variables associated with nodes in a single interconnected network and with dependence structure informed by the network ties. We demonstrate that Moran's I and Φ provide valid tests for such dependence.

For each of four simulation settings we generated a fully connected social network with $n = 200$ nodes. We simulated i.i.d., mean-zero starting values for each node and then ran several iterations of a direct transmission process, by which each node is influenced by its neighbors, to generate a vector of outcomes $\mathbf{Y} = (Y_1, Y_2, ..., Y_{200})$ associated with the nodes. We ran the simulation 500 times for each setting, generating 500 outcome vectors. While the amount of network

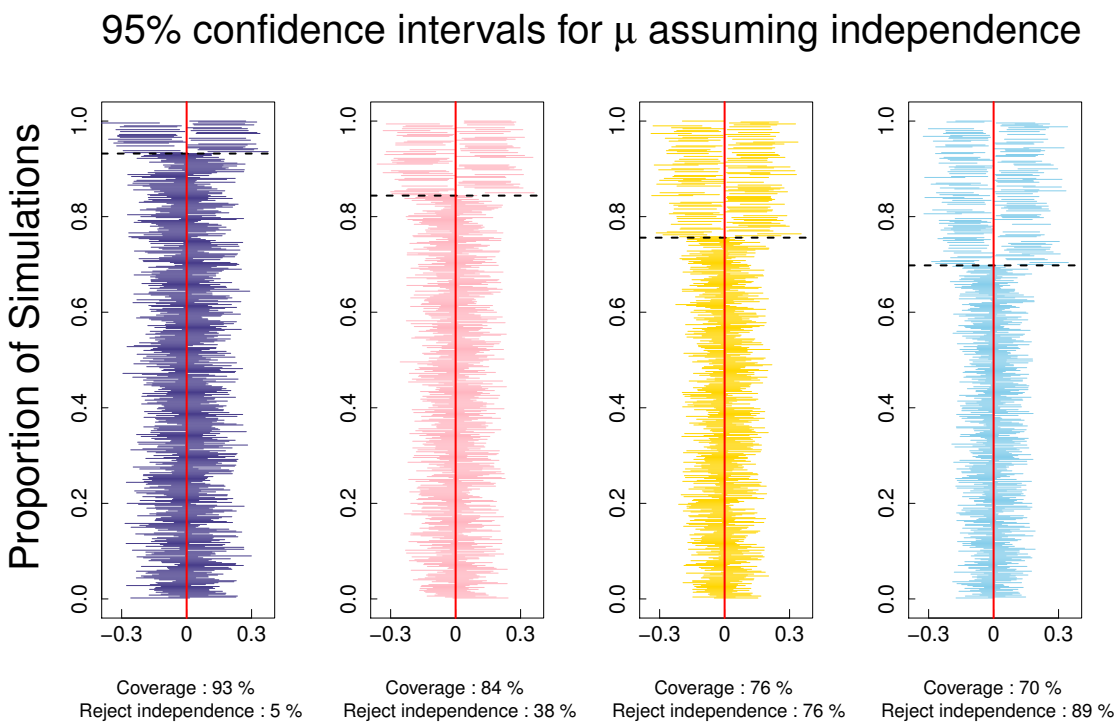


Figure 2.3: Each column contains 95% confidence intervals (CIs) for $E[Y] = \mu$ under dependence due to direct transmission, with increasing dependence from left (no dependence) to right. The CIs above the dotted line do not contain the true $\mu = 0$ (red-line) while the CIs below the dotted line contain μ. Coverage rates of 95% CIs are calculated as the percentages of the CIs covering μ. We also present the percentages of permutation tests based on Moran's I that reject the null at $\alpha = 0.05$; this is the type I error for the leftmost column and the power for the other three columns.

dependence in the outcomes varied across simulation settings (controlled by the number of iterations of the spreading process), the expected outcome $E[Y]$ was 0 for every setting. To demonstrate the impact of using i.i.d. methods when dependence is present, in each simulation we calculated a 95% confidence interval (CI) for $E[Y]$ under the assumption of independence. We estimated the mean of $E[Y]$ using $\bar{Y}$ and we estimated the standard error (s.e.) for $\bar{Y}$ under

the assumption of independence, that is ignoring the presence of any pairwise covariance terms. The 95% confidence interval is given by $\bar{Y} \pm 1.96 * s.e.$ In each simulation we also ran a test for network dependence using Moran's I.

Figure 2.3 displays the results of four simulation settings, with increasing dependence from left to right. The left-most column represents a setting with no dependence. Each column depicts 500 95% confidence intervals, one for each simulation. The confidence intervals below the dotted lines cover the true mean of 0, while the intervals above the dotted line do not. The coverage is close to the nominal 95% under independence, but decreases dramatically as dependence increases, despite the fact that $\bar{Y}$ remains unbiased for $E[Y]$. We also report the power of permutation tests based on Moran's I (with subject index randomly permuted $M = 500$ times) to reject the null hypothesis of independence at the $\alpha = 0.05$ level. Under independence the test rejects 5% of the time, as is to be expected, and as dependence increases and coverage decreases, the power of our test to detect dependence increases, achieving almost 90% when the coverage drops below 70%. (That the power to detect dependence increases with increasing dependence is robust to the specifics of the simulations, but the exact relation between coverage and power is not; in other settings 90% power could correspond to different coverage rates.) These results highlight the fact that a strict $p < 0.05$ cut-off may not be appropriate for these tests of dependence.

Table 2.1: Coverage rate of simultaneous 95% CIs, empirical power of tests of independence using asymptotic normality of Φ, and empirical power of permutation tests of independence based on Φ, under direct transmission for $t = 0, 1, 2, 3$. The size of the tests is $\alpha = 0.05$.

	95% CI coverage rate	% of p-values(z) ≤ 0.05	% of p-values(permutation) ≤ 0.05
t=0	0.94	5.40	4.80
t=1	0.81	39.40	36.20
t=2	0.63	67.80	65.00
t=3	0.43	85.40	83.40

To illustrate the performance of Φ, we simulate a categorical outcome Y with five levels and with marginal probabilities $(p_1,\ p_2,\ p_3,\ p_4,\ p_5) = (0.1,\ 0.2,\ 0.3,\ 0.25,\ 0.15)$. To demonstrate the consequences of using i.i.d. inference in the presence of dependence, we calculated simultaneous 95% confidence intervals for estimates of p_1 through p_5 using the method of (Sison and Glaz, 1995). We also report the power to reject the null hypothesis of independence as the percentage of 500 simulations in which hypothesis tests based on our new statistic, Φ, rejected the null. Table 2.1 summarizes the simulation results for dependence by direct transmission. It is evident that as dependence increases, coverage rates of i.i.d. 95% confidence intervals decrease, and the power to reject the null increases. Details of the simulation models and results from additional simulations are provided in the Supplementary Materials. The R function for testing network dependence and generating network dependent observations can be found in the netdep R package available at Github (`github.com/youjin1207/netdep`).

2.4 Applications

2.4.1 Spatial data

In this section we apply Φ to spatial data on 473 power producing facilities that we introduced in Section 2.3.1, and compare the results to standard analyses using join count statistics. In addition to the locations of the 473 facilities, the data includes information on the characteristics of the surrounding geographic areas. Details can be found in Table G.1 in Papadogeorgou et al. (2016).

In Figure 2.4a, we mapped the proportion of the populations within a 100km radius around each of the facilities falling into three different race/ethnicity categories. We can apply Moran's I separately to each of the three proportions, but Moran's I cannot provide a single aggregate test statistic aggregating the three proportions. For example, we may be interested in autocorrelation with respect to the dominant demographic group (Table 2.2) or regions with more than 10% Hispanics and African Americans (Table 2.3). Tables 2.2 and 2.3 respectively present the frequency of concordant neighboring pairs with these characteristics, and the corresponding join count analysis results. To calculate the join count statistics, we specify a neighborhood size of 15, meaning that observation j is considered to be adjacent to i if j is one of i's closet 15 neighbors in Euclidean distance.

Table 2.2: Permutation tests of dependence based on join count statistics applied to dominant race/ethnicity group.

Dominant group	White	Hispanic	African-American
n	446	13	14
Join count statistic	212.63	0.97	0.77
P-value (permutation)	0.0020	0.0020	0.0020

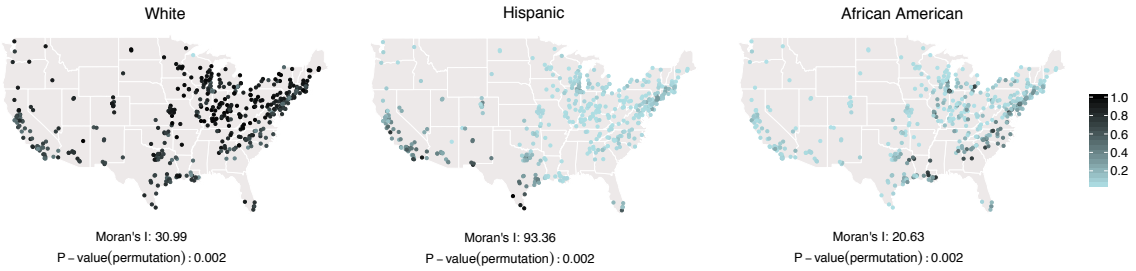
Table 2.3: Permutation tests of dependence based on join count statistics applied to four different population categories, defined by having $\leq$10% or $>$10% Hispanic or African American residents.

	AA $>$ 10%, HP $>$ 10%	AA $>$ 10%, HP $\leq$ 10%	AA $\leq$ 10%, HP $>$ 10%	AA $\leq$ 10%, HP $\leq$ 10%
n	52	106	98	217
Join-count statistic	7.07	26.63	30.30	69.20
P-value (permutation)	0.0020	0.0020	0.0020	0.0020

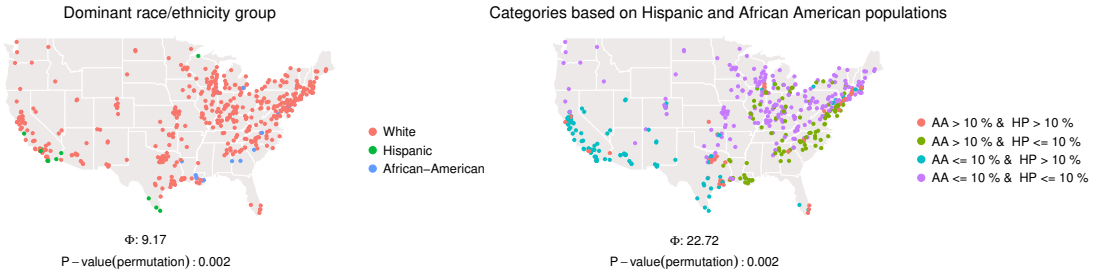
In Figure 2.4b, we map the distribution of dominant racial group and regions with more than 10% Hispanics and African Americans and give an omnibus test for autocorrelation based on Φ. We observe higher autocorrelation in the second categorization ($\Phi : 22.72$) than the first categorization ($\Phi : 9.17$), which cannot be compared from join count statistics presented in Table 2.2 and Table 2.3.

2.4.2 Network data

The Framingham Heart Study, initiated in 1948, is an ongoing cohort study of participants from the town of Framingham, Massachusetts that was originally designed to identify risk factors for cardiovascular disease. The study has grown over the years to include five cohorts. The original cohort ($n = 5,209$) was originally recruited in 1948 and has been continuously followed since then.



(a) Proportion of race/ethnicity groups around 473 power-producing facilities across the U.S.. Applying Moran's I separately to each proportion, all of the tests reject the null hypothesis of independence at the $\alpha = 0.05$ level.



(b) Dominant group (left) and categories defined by having $\leq$10% or $>$10% Hispanic or African American residents (right). Omnibus tests of dependence based on Φ reject the null hypothesis of independence at the $\alpha = 0.05$ level for both variables.

Figure 2.4

The offspring cohort ($n = 5,124$) was initiated in 1971 and includes offspring of the original cohort members and the offspring's spouses. The third generation cohort ($n = 4,095$), initiated in 2001, is comprised of offspring of members of the offspring cohort. Spouses of members of the offspring cohort who were not themselves included in that cohort and whose children had been recruited into the third generation cohort were invited to join the New Offspring Spouse Cohort ($n = 103$) beginning in 2003. Two omni cohorts (combined $n = 916$) were started in 1994 and 2003 in order to reflect the increasingly diverse population of Framingham; these cohorts specifically targeted residents of Hispanic,

Asian, Indian, African American, Pacific, Islander and Native American descent.

Members of the original cohort are followed through biennial examinations while members of other cohorts are examined every 4 to 8 years. Each examination includes non-invasive tests, e.g. X-ray, ECG tracings, or MRI; laboratory tests of blood and urine; questionnaires pertaining to diet, sleep patterns, physical activities, and neuropsychological assessment; and a physical exam, including assessments for cardiovascular disease, rheumatic heart disease, dementia, atrial fibrillation, diabetes, and stroke. Other measures and tests are collected sporadically. In addition, in between each exams, participants are regularly monitored through phone calls. Genotype and pedigree data has been collected for all (consenting) participants, and the study populations includes multiple members of 1538 families, making the FHS a powerful resources for heritability studies. Detailed information on data collected in the FHS can be found in Tsao and Vasan (2015). Public versions of FHS data from the original, offspring, new offspring spouse, and generation 3 cohorts through 2008 are available from the dbGaP database.

For decades, FHS has been one of the most successful and influential epidemiologic cohort studies in existence. It is arguably the most important source of data on cardiovascular epidemiology. It has been analyzed using i.i.d. statistical models (as is standard practice for cohort studies) in over 3,400 peer-

reviewed publications since 1950: to study cardiovascular disease etiology (e.g. Castelli 1988; D'Agostino et al. 2000, 2008), risks for developing obesity (e.g. Vasan et al. 2005), factors affecting mental health (e.g. Qiu et al. 2010; Saczynski et al. 2010), cognitive functioning (e.g. Au et al. 2006), and many other outcomes.

In addition to being a very prominent cohort study, the FHS plays a uniquely influential role in the study of social networks and social contagion. Leading up to the publication of Christakis and Fowler (2007), researchers discovered an untapped resource buried in the FHS data collection tracking sheets: information on social ties that allowed them to reconstruct the (partial) social network underlying the cohort. The tracking sheets were originally intended to facilitate exam scheduling, and they asked each participant to name close contacts who could help researchers to locate the participant if the participant's contact information changed. Combining this information with existing data on family and spousal connections, researchers were able to build a partial social network with ties representing friends, co-workers, and relatives. They then leveraged this social network data to study peer effects for obesity (Christakis and Fowler, 2007), smoking (Christakis and Fowler, 2008), and happiness (Fowler and Christakis, 2008). The FHS has since been used to study peer effects by many other researchers (Pachucki et al., 2011; Rosenquist et al., 2010).

We analyzed data from the Offspring Cohort at Exam 5, which was con-

ducted between 1991 and 1995. Because the publicly available data are divided into datasets for individuals with and without non-profit use (NPU) consent and these two datasets have separate network data, we only used data from the NPU consent group, giving us a sample size of 1,033 with 690 undirected social network ties.

Figure 2.5 depicts the distribution of systolic and diastolic blood pressure over the five largest connected network components; darker colors represent higher blood pressure values. We used Moran's I to test for network dependence in these two continuous random variables. – systolic blood pressure and diastolic blood pressure. We found significant evidence of network dependence in systolic blood pressure (p-value : 0.03), but not for diastolic blood pressure (p-value : -0.87).

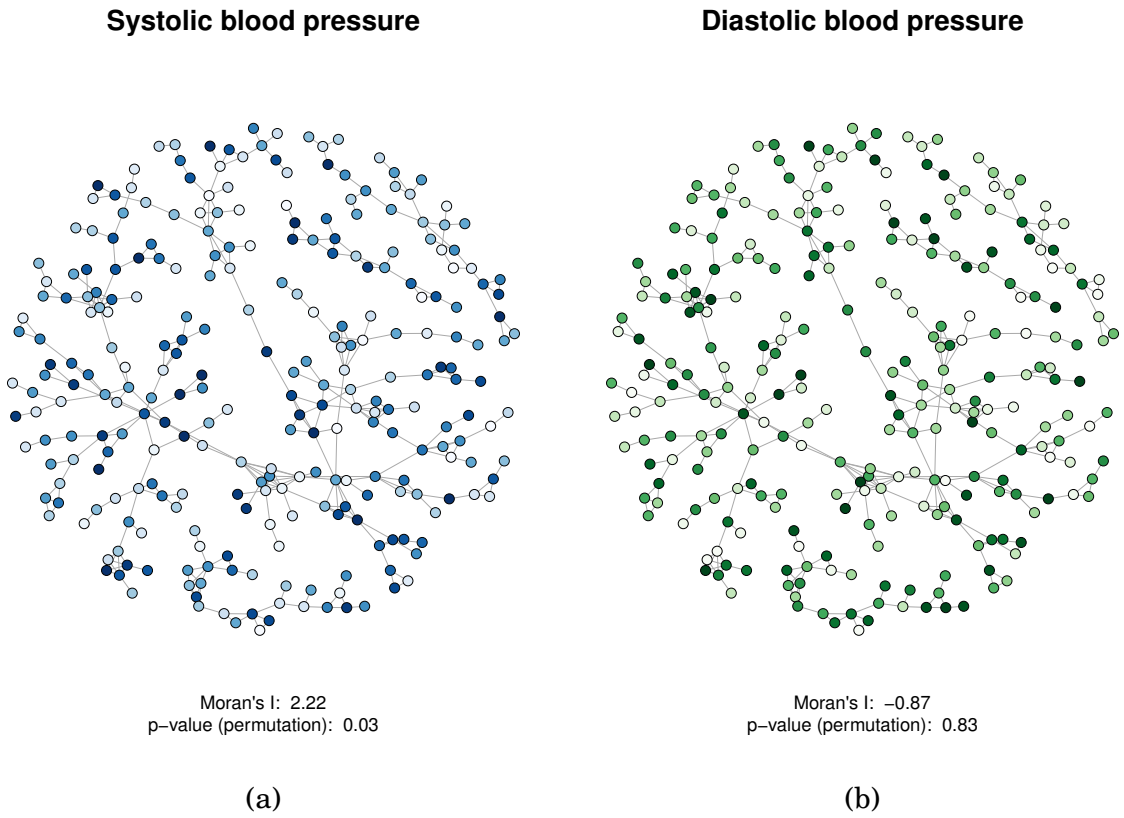


Figure 2.5: The five largest connected components, encompassing 273 subjects, in the social network of 1,031 subjects from the FHS Offspring Cohort Exam 5 data. The color of the node represents the subject's blood pressure values: high values of systolic blood pressure and diastolic blood pressure are darker and low values are lighter.

We tested for dependence in two different categorical random variables using Φ: employment status and preferred method of making coffee. Figure 2.6 shows the distribution of the two variables over the largest connected component of the network. We found significant evidence of network dependence for both variables.

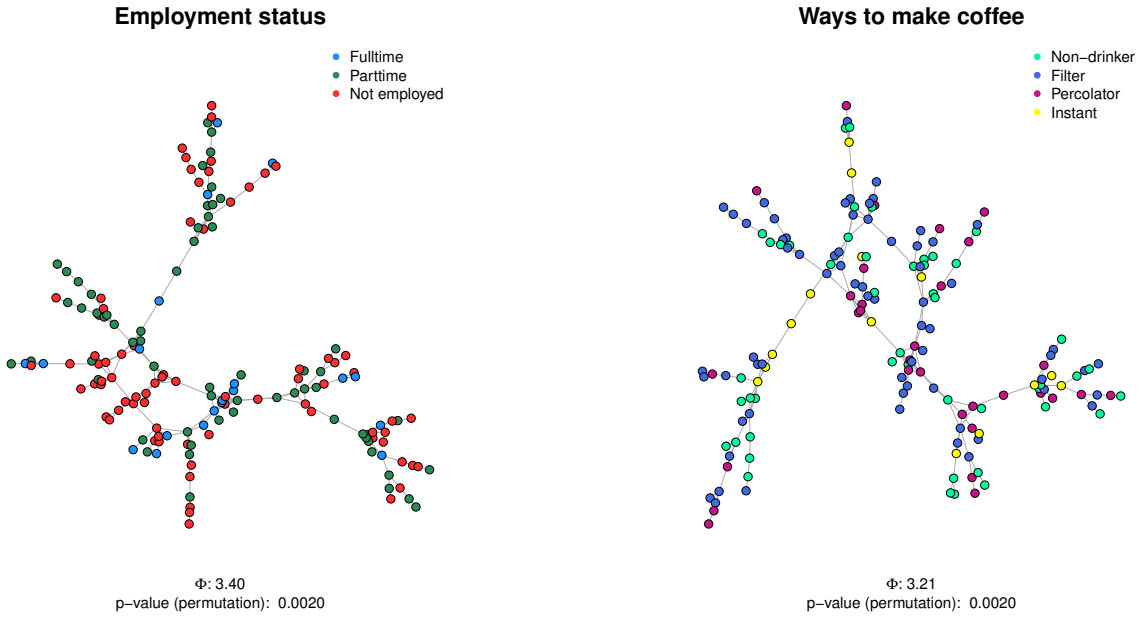


Figure 2.6

2.5 Concluding Remarks

In this chapter, we proposed simple tests for independence among observations sampled from geographic space or from a network. We demonstrated the performance of our proposed tests in simulations under both spatial and network dependence, and applied them to spatial data on U.S. power producing facilities and to social network data from the Framingham Heart Study.

Under network dependence, adjacent pairs are expected to exhibit the greatest correlations, and for robustness we used the adjacency matrix as the weight matrix for calculating the test statistic, thereby restricting our analysis to adja-

cent pairs; if researchers have substantive knowledge of the dependence mechanism other weights may increase power and efficiency.

Researchers should be aware of the possibility of dependence in their observations, both when studying social networks explicitly and when observations are sampled from a single community for reasons of convenience. As we have seen in the classic Framingham Heart Study example, such observations can be correlated, potentially rendering i.i.d. statistical methods invalid. In a forthcoming companion paper, we illustrate the consequences of assuming that observations are independent when they may in fact exhibit network dependence.

Acknowledgments

Youjin Lee and Elizabeth Ogburn were supported by ONR grant N000141512343. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195 and HHSN268201500001I). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

2.6 Appendix

2.6.1 Moments of Φ

Let $\mu_\Phi := E[\Phi]$ and $E[\Phi^2]$ be the first and second moments of Φ respectively. Based on these moments, we can derive the variance of Φ, $\sigma^2_\Phi := E[\Phi^2] - \mu^2_\Phi$.

$$
\begin{aligned}
\mu_\Phi &= \frac{1}{n(n-1)}\{n^2k(2-k) - nQ_1\} \\
E[\Phi^2] &= \frac{1}{S_0^2}\Bigg[\frac{S_1}{n(n-1)}(n^2Q_{22} - nQ_3) \\
&\quad + \frac{S_2 - 2S_1}{n(n-1)(n-2)}((k-4)k+4)n^3Q_1 + n(n((2k-4)Q_2 - Q_{22}) + 2Q_3) \\
&\quad + \frac{S_0^2 - S_2 + S_1}{n(n-1)(n-2)(n-3)}\Big\{n(-4Q_3 + 2nQ_{22} - 6knQ_2 + 12nQ_2 \\
&\quad - 3k^2n^2Q_1 + 14kn^2Q_1 - 16n^2Q_1 + k^4n^3 - 4k^3n^3 + 4k^2n^3) \\
&\quad - ((2k-4)n^2Q_2 + n^2(kn(2Q_1 - kQ_1) - Q_{22}) + 2nQ_3)\Big\}\Bigg],
\end{aligned}
\tag{2.4}
$$

where k is the number of categories; $Q_m := \sum_{l=1}^{k} 1/p_l^m, (m = 1,2,3)$; $Q_{22} := \sum_{l=1}^{k}\sum_{u=1}^{k} 1/p_l p_u$; $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij}+w_{ji})/2$; $S_1 = \sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij}+w_{ji})^2/2$; $S_2 = \sum_{i=1}^{n}(w_{i\cdot}+w_{\cdot i})^2$.

2.6.2 Asymptotic Distribution of Φ under the Null

Shapiro and Hubert (Shapiro and Hubert, 1979) proved the asymptotic normality of permutation statistics of the form H_n for *i.i.d* random variables $Y_1, Y_2, \ldots, Y_n$

under some conditions:

$$H_n = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} d_{ij} h(Y_i, Y_j), \tag{2.5}$$

where $h(\cdot,\cdot)$ is a symmetric real valued function with $E[h^2(Y_i, Y_j)] < \infty$ and $\mathbf{D} := \{d_{ij}; i, j = 1, ..., n\}$ is a $n \times n$ symmetric, nonzero matrix of which all diagonal terms must be zero. In the context of Φ, $h(Y_i, Y_j) = \big(2I(Y_i = Y_j) - 1\big)/(p_{Y_i} p_{Y_j})$ and $\mathbf{D} = \mathbf{W}$. Requirements for asymptotic normality include $\sum\limits_{i,j=1,j\neq i}^{n} d_{ij}^2 / \sum\limits_{i=1}^{n} d_{i\cdot}^2 \to 0$ and $\max\limits_{1\leq i\leq n} d_{i\cdot}^2 / \sum\limits_{k=1}^{n} d_{k\cdot}^2 \to 0$ as $n \to 0$ for $d_{i\cdot} = \sum\limits_{j=1}^{n} d_{ij}$. If we use the adjacency matrix for $\mathbf{W}$, this implies $\sum\limits_{i,j=1,i\neq j}^{n} A_{ij} / \sum\limits_{i=1}^{n} A_{i\cdot}^2 \to 0$ and $\max\limits_{1\leq i\leq n} A_{i\cdot} / \sum\limits_{i=1}^{n} A_{i\cdot}^2 \to 0$ where $A_{i\cdot}$ is the degree of node i. More details can be found in Shapiro and Hubert (1979); see also O'Neil and Redner (1993).

Chapter 3

Invalid Statistical Inference Due to Social Network Dependence

Researchers across the health and social sciences generally assume that observations are independent, but when observations are dependent, using statistical methods that assume independence can lead to biased estimates (with bias away from the null) and to artificially small p-values, standard errors, and confidence intervals. This results in inflated false positive rates and may contribute to replication crises. Here, we describe a largely unrecognized but common type of dependence due to social network connections, and explain how such dependence increases variance and engenders confounding that can lead to biased estimates. We describe network dependence and introduce the concept of *confounding by network structure*. We apply a test for network de-

pendence to several published papers that use the Framingham Heart Study (FHS) data. Results suggest that some of the many decades of research on coronary heart disease, other health outcomes, and peer influence using FHS data may be invalid due to unacknowledged network dependence. The FHS is not unique; these problems could arise whenever human subjects are recruited from one or a small number of communities, schools, hospitals, etc. As researchers in psychology, medicine, and beyond grapple with replication failures, this unacknowledged source of invalid statistical inference should be part of the conversation.

This is a joint work in collaboration with Elizabeth Ogburn.

3.1 Introduction

The replication crises in psychology, medicine, and other fields have drawn attention to many ways that the flawed application of statistics can result in spurious findings. In this paper we identify an unacknowledged but potentially pervasive source of invalid statistical inference that could lead to inflated false positive rates, namely social network dependence.

Assuming that data are independent and identically distributed (i.i.d.) is the default for most applications of statistics, but when i.i.d. statistical methods are used to analyze data that are in fact dependent, the resulting infer-

ence is generally anticonservative: standard errors, p-values, and confidence intervals are artificially small. This can lead to inflated false positive rates. Whenever human subjects are sampled from one or a small number of communities, schools, hospitals, etc., as is routine in the health and social sciences, they may be connected by social ties, such as friendship or family membership, that could engender statistical dependence, which we refer to as *network dependence*. When an outcome and an exposure of interest both exhibit network dependence, estimates of associations will often be biased away from the null due to *confounding by network structure*. Yet the i.i.d. assumption is seldom questioned or tested, and the possible presence of social network dependence is routinely ignored even when subjects are recruited from a single close-knit community, as in the influential Framingham Heart Study, which we use to illustrate these problems.

We define *network dependence* and *confounding by network structure*, describe tests that can help detect when these might be a problem in real data, and illustrate how ignoring these features of data can result in biased and invalid statistical inference. We test for network dependence and for possible confounding by network structure in several published analyses using data from the Framingham Heart Study (FHS), which is a paradigmatic example of an epidemiologic study comprised of individuals who are all members of a single tight-knit community. The FHS data includes some explicit informa-

tion about network ties, and researchers have used these data to study social network phenomena such as social contagion, also with i.i.d. methods. Our results suggest that the i.i.d. assumption—on which thousands of FHS papers have relied—does not reliably hold, and that confounding by network structure may be widespread.

3.2 Network Dependence

A network is a collection of nodes and edges (Newman, 2010), where, in a social network, a node represents a person and an edge connecting two nodes represents the existence of some relationship or social tie between them. When the nodes in a network correspond to students in a high school, for example, a tie may indicate that two students are in the same class or that they are members of the same school club; when nodes are patients staying in a hospital, a tie between patients may represent a shared doctor or a shared hospital unit.

In the literature on spatial and temporal dependence, dependence is often implicitly assumed to be the result of latent traits that are more similar for observations that are close than for distant observations. This latent variable dependence (Ogburn, 2017) is likely to be present in many network contexts as well. Homophily, or the tendency of similar people to form network ties, is a paradigmatic source of latent trait dependence. If the outcome under study in

a social network has a genetic component, then we would expect latent variable dependence due the fact that family members, who share latent genetic traits, are more likely to be close in social distance than people who are unrelated. If the outcome is affected by geography or physical environment, latent variable dependence could arise because people who live close to one another are more likely to be friends than those who are geographically distant. In networks, edges often present opportunities to transmit traits or information from one node to another, and such direct transmission will result in dependence that is informed by the underlying network structure (Ogburn, 2017). In general, both of these sources of dependence result in positive pairwise correlations that tend to be larger for pairs of observations from nodes that are close in the network and smaller for observations from nodes that are distant in the network.

To illustrate the consequences of treating network observations as if they are i.i.d., consider a hypothetical sample of n nodes in a social network, e.g. students at a U.S. college with ties representing friendship, cohabitation, participation in the same activities, etc.. Each node provides an outcome Y, e.g. body mass index (BMI). Suppose that, as has been suggested by some researchers (Christakis and Fowler, 2007), BMI exhibits network dependence due to "social contagion." The target of inference is the mean μ of BMI for U.S. college students. The sample average $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ is unbiased for μ as long as the students at this particular college are representative of the overall U.S. col-

lege student population. While bias and representability are not necessarily affected by social network connections, the variance of $\bar{Y}$ will be affected by network dependence. For the purposes of this example, suppose that $Y_1, Y_2, ..., Y_n$ are identically but not independently distributed, with common mean μ and variance σ^2. Then

$$\begin{aligned} Var(\bar{Y}) &= Var\left(\sum_{i=1}^{n} Y_i\right)/n^2 \\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\sigma^2 + \sum_{i\neq j} cov(Y_i, Y_j)\right\} \\ &= \frac{\sigma^2}{n/\left(1+\frac{b_n}{\sigma^2}\right)}, \end{aligned} \tag{3.1}$$

where $b_n = \frac{1}{n}\sum_{i\neq j}^{n} cov(Y_i, Y_j)$. The quantity $n/\left(1+\frac{b_n}{\sigma^2}\right)$ in the denominator is the *effective sample size* of the dependent sample, and under dependence it is generally smaller than the apparent sample size n. But it is the effective rather than the apparent sample size that determines standard errors and rates of convergence for dependent samples. A researcher who failed to question the independence of $Y_1, Y_2, ..., Y_n$ would estimate $Var(\bar{Y})$ with σ^2/n, but whenever b_n is positive (as is expected under network dependence), this underestimates the true variance. Inference using variance estimators based on σ^2/n will be anticonservative: p-values will be artificially low and confidence intervals artificially narrow. With more dependence b_n increases, the effective sample size decreases, and inference that assumes independence is more anticonservative.

Very informally, when subjects are independent, each new observation brings one new ”bit” of information about μ; when subjects are dependent, each new observation brings less than one new ”bit” of information because some of the information is redundant due to dependence on the previous observations. Therefore, a researcher who falsely assumes independence believes that the data provide more information than they actually do, i.e. the researcher overestimates the strength of evidence provided by the data.

In some settings researchers routinely account for statistical dependence in data analyses: for example, when data are clustered (e.g. clustered randomized trials, batch effects in lab experiments), when studying genetics or heritability in a sample of genetically related organisms, or when data may exhibit spatial or temporal dependence. But outside of these settings it is generally standard practice to use statistical methods that assume independent and identically distributed (i.i.d.) data. Despite increasing interest in and availability of social network data, there is a dearth of valid statistical methods to detect or account for network dependence.

3.2.1 Regression models

Coefficients from regression models suffer from the same problems as sample means in the presence of network dependence. Standard regression models assume independent errors, but when an outcome exhibits network de-

pendence the regression errors generally will, too, rendering inferences drawn from the regression models invalid. Although researchers have developed regression models for many kinds of dependent data, it is not clear that any of them are generally appropriate for social network data, and certainly none are in wide use for network data.

3.2.2 Confounding by network structure

Bias can result when both an outcome and a covariate of interest exhibit network dependence. In this case, the network structure can act like a confounder, creating a spurious association between the covariate and outcome. Returning to the example above, suppose researchers use data from the college students to ascertain whether choice of academic major is associated with BMI. Students form strong friendships with other students having similar academic interests, engendering network dependence in academic major. An entirely independent process engenders network dependence in BMI: obesity is socially contagious, so students who are friends with one another (regardless of whether the friendship is related to shared academic interests) tend to have similar BMI. Due solely to the underlying network structure, students with the same major are expected to have similar BMI. We would not expect to see this same association in an i.i.d. sample, for example a national sample drawing independent students from many different colleges. Confounding by network

structure is analogous to confounding by population stratification and confounding by cryptic relatedness, two well-known sources of bias in population-based genetic association studies when both the outcome and the (in this case genetic or genomic) covariate of interest share a common dependence structure (Sillanpää, 2011).

3.2.3 Testing for network dependence

In a companion technical report (Lee and Ogburn, 2018b) we propose statistical methods to test for the presence of network dependence in data with some information about network ties, based on Moran's I, a well-known statistic from the spatial autocorrelation literature. An R package is available (Lee and Ogburn, 2018a). The test takes as inputs a single value associated with each subject, e.g. an outcome, predictor, or regression residual, and a weighted distance matrix with an entry for each pair of subjects. The weight matrix should place higher weights on pairs of subjects who are close in network distance and smaller weights on pairs of subjects who are distant in the network. The choice of weights affects the power, but not the validity, of the test. Similarly, if information is available about some but not all network ties, this will tend to reduce the power of the test but not affect its validity. A robust choice of weight matrix is the *adjacency matrix* for the network, which puts weight 1 on pairs of subjects who share a network tie and weight 0 otherwise; we use this

weight matrix throughout. We recommend viewing moderate to large statistics as evidence of possible dependence even if p-values do not meet an arbitrary $\alpha = 0.05$ cut-off, and caution that network dependence may be present even if these statistics are small. If the test statistic calculated from regression residuals is moderate to large, it suggests that standard error estimates from i.i.d. regression models may be underestimated. If both of the test statistics calculated from an outcome and from a covariate of interest are moderate to large, it suggests that confounding by network structure may be present.

3.3 Framingham Heart Study

The Framingham Heart Study (FHS), initiated in 1948, is arguably the most important source of data on cardiovascular epidemiology. It is also an influential source of data on network peer effects. FHS is an ongoing cohort study of participants from the town of Framingham, Massachusetts, that has grown over the years to include five cohorts with a total sample of over $15,000$, representing almost 25% of the total population of Framingham. Multiple members (> 3) of more than $1,500$ extended families are included in the study population. Study participants are followed through exams every 2 to 8 years. In between exams, participants are regularly monitored through phone calls. Detailed information on data collected in the FHS can be found in Tsao and Vasan

(2015). Public versions of FHS data through 2008 are available from the dbGaP database. The FHS data have been analyzed using i.i.d. statistical models (as is standard practice for cohort studies) in over 3,400 peer-reviewed publications since 1950, most of which use multiple regression to explore associations between cardiovascular outcomes and various risk factors. Because the individuals in the FHS are members of a single community, connected by social and familial ties, the outcomes and covariates of interest may be exhibit network dependence. Yet to our knowledge, none of the published studies using FHS data has acknowledged this possibility, including in the literature on peer effects.

Below we demonstrate the potential for bias due to confounding by network structure and show that there is evidence of potentially widespread dependence in the outcomes, predictors, and regression residuals from published papers using FHS data. The problem of network dependence extends to high profile research using FHS data to explicitly study peer effects and social contagion in social networks, but with statistical methods designed for i.i.d. data.

3.3.1 Confounding by network structure

In order to demonstrate the bias that can arise when both a predictor and an outcome share common network structure, we simulated a covariate with dependence structure governed by the FHS social network but otherwise unre-

lated to any of the variables measured in the FHS. We generated a continuous network dependent covariate, X, conditional on the FHS network, independently 500 times. We regressed a cardiovascular outcome (systolic blood pressure, SBP), a lifestyle outcome (employed or not), a health-seeking behavior outcome (visited a doctor due to illness), and a non-cardiovascular health outcome (diagnosis of corneal arcus) from the FHS data onto X. For each of the four outcomes we fit the same regression model independently 500 times, once for each of the independently generated covariates.

Figure 3.1 shows the coverage of 95% confidence intervals for β, the coefficient for X in the regression of each outcome onto X plus an intercept. Because the covariate is generated without reference to any of these outcomes, the true value of β for a population-based, rather than network, sample is 0. However, for all four outcomes the confidence intervals are not centered around 0, indicating that estimates of β are biased due to confounding by network structure. For all four outcomes the confidence intervals exhibit undercoverage, ranging from 65% to 85% rather than the nominal rate of 95%. While the bias is due to confounding by network structure; the undercoverage may be due to both confounding and to network dependence in the regression residuals, which could result in underestimated standard errors. Table 3.1 reports the p-values for tests of dependence in the four outcomes, the predictor X (averaged across 500 replicates), and the residuals from the regression of the outcome on X (aver-

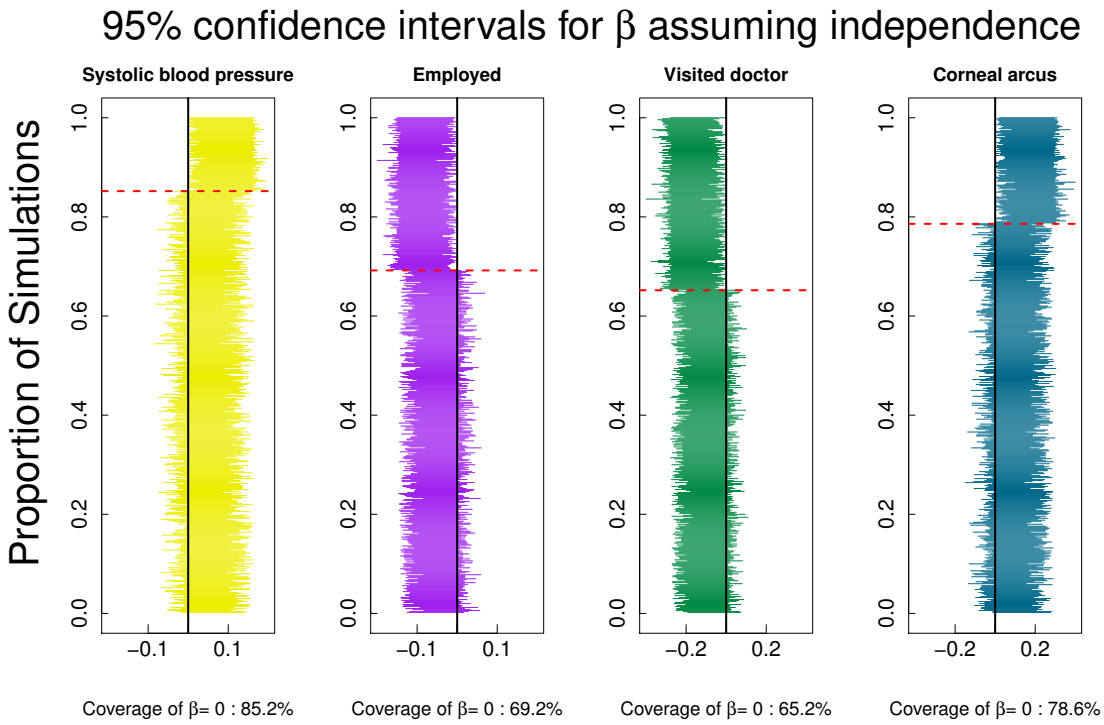


Figure 3.1: Each column contains 95% confidence intervals (CIs) for the coefficient for a random, network dependent covariate. The CIs above the dotted line do not contain the null value $\beta = 0$ (red-line) while the CIs below the dotted line contain 0. Coverage rates of 95% CIs are calculated as the percentages of the CIs covering 0.

aged across 500 replicates for each outcome). For three of the outcomes (SBP, employment, and corneal arcus) tests based on Moran's I suggested strong evidence of dependence; for visit to doctor the test did not show strong evidence of dependence in the outcome or residuals (though we reiterate that a null test does not imply a lack of dependence). Simulation and analysis details are in the Supplementary Materials.

Table 3.1: Results of tests of network dependence for the outcomes, simulated predictor X, and residuals from regressing each outcome onto X. P-values are obtained from permutation tests.

	Systolic blood pressure	Employed	Visited doctor	Corneal arcus
p-value for outcome	0.03	0.00	0.71	0.01
Average p-value for predictor	0.00	0.00	0.00	0.00
Average p-value for residuals	0.04	0.00	0.70	0.02

3.3.2 Cardiovascular disease epidemiology

In order to evaluate whether network dependence and confounding due to network structure may undermine research using FHS data, we chose regression models from five published papers in the epidemiologic and medical literature and applied our tests of dependence to the outcomes, covariates, and regression residuals. We screened for ease of replicability using publicly available data (i.e. models are explicitly defined using variables that are available in the public data), and selected the first five papers that we found on Google Scholar that met the replicability criteria. Because we require social network information for our tests of dependence, and because that information is not available for all individuals and is not straightforward to harmonize across exams, we ran the published regression models on subsets of the data for which network information was readily available. Below we report results from the two papers for which we found the strongest evidence of dependence: the models reported in these two papers show compelling evidence of network dependent outcomes, covariates, and residuals. We also found moderate evidence of

dependence in some of the analyses reported in each of the other three papers (Wolf et al., 1991; Gordon et al., 1977; Levy et al., 1990); details are in the Supplementary Information.

Lauer et al. (1991) examined the association between obesity and left ventricular mass (LVM); this paper is one of the authors' many highly cited papers on LVM, which is of interest to many researchers due to its relationship with cardiovascular disease (Levy et al., 1990) and other cardiovascular outcomes. The study assessed the relationship between obesity and LVM using the estimated coefficients for BMI in sex-specific linear regressions adjusted for age and systolic blood pressure, where the outcome was LVM normalized by height. This analysis indicated that obesity is a significant predictor of LVM conditional on age and systolic blood pressure for both men and women.

In order to test whether the assumptions of independence inherently assumed by Lauer et al. (1991) are valid, we applied Moran's I to normalized LVM and to BMI, separately for males and females, and to the residuals from our replication of the Lauer et al. sex-specific regressions. The results are reported in Table 3.2. In order for the inference reported in Lauer et al. (1991) to be valid, the errors from the regressions should be independent, however Moran's I provides evidence of network dependence for the residuals in addition to the marginal LVM variable, for both males and females, undermining the i.i.d. assumption on which the validity of the linear regression model rests.

Table 3.2: Results of tests of network dependence for males and females, for LVM, BMI, and the residuals from regressing LVM onto covariates. P-values are obtained from permutation tests.

Y	I_{std}	P-value
Male		
Normalized LVM	2.26	0.01
BMI	1.36	0.09
Residual from LVM ~ BMI + age + systolic BP	1.34	0.11
Female		
Normalized LVM	2.23	0.02
BMI	1.51	0.06
Residual from LVM ~ BMI + age + systolic BP	2.92	0.00

Furthermore, for both sexes there is evidence of network dependence for both LVM and BMI, suggesting that any association may be due to confounding by network structure.

Cox proportional hazards models (Cox, 1992) are commonly applied to the FHS data to assess risk factors for mortality. When the assumptions of the Cox model hold, including i.i.d. observations, Martingale residuals are expected to be approximately uncorrelated in finite samples (Lin et al., 1993; Tableman and Kim, 2003). We looked for evidence of residual dependence in a study by Tsuji et al. (Tsuji et al., 1994) of the association between eight different heart rate variability (HRV) measures and four-year mortality. We replicated the twenty-four separate Cox models reported in Tsuji et al. (1994): for each of eight measures of HRV we fit models without adjusting for covariates, adjusting for age and sex, and adjusting for clinical risk factors in addition to age and sex.

Table 3.3 shows the results of applying tests of independence using Moran's

Table 3.3: Tests of network dependence using Moran's I statistic applied to each HRV measure and to the Martingale residuals from the Cox models for eight different HRV measures. P-values are obtained from permutation tests.

HRV measures:	lnSDNN	lnpNN50	lnr-MSSD	lnVLF	lnLF	lnHF	lnTP	lnLF/HF
Covariate								
I_{std}	0.33	-0.41	-0.12	1.72	1.62	0.83	1.85	-0.03
P-value	0.38	0.59	0.52	0.06	0.08	0.20	0.06	0.47
Residuals from unadjusted model for all-cause mortality								
I_{std}	1.57	1.65	1.64	1.38	1.38	1.54	1.38	1.59
P-value	0.06	0.04	0.04	0.08	0.09	0.06	0.08	0.05
Residuals from model for all-cause mortality adjusted for age and sex								
I_{std}	1.94	2.00	2.05	1.92	1.75	1.95	1.87	1.97
P-value	0.02	0.02	0.02	0.02	0.04	0.02	0.03	0.03
Residuals from model for all-cause mortality adjusted for age, sex, and clinical risk factors								
I_{std}	1.55	1.52	1.56	1.60	1.46	1.53	1.52	1.52
P-value	0.07	0.07	0.07	0.06	0.09	0.07	0.09	0.07

I to the Martingale residuals from the twenty-four different regression models, which suggest that the i.i.d. assumption may be violated in most or all of these regressions. Interestingly, Moran's I statistic is larger with smaller p-values for the covariates that were found to be significant predictors of all cause mortality. This is consistent with a hypothesis that the statistically significant associations are due to confounding by network structure rather than to true population-level associations.

3.3.3 Peer effects

The FHS plays a uniquely influential role in the study of social networks and social contagion. Christakis and Fowler (C&F) discovered an untapped resource buried in the FHS data collection tracking sheets: information on social ties that, combined with existing data on connections among the FHS partici-

pants, allowed them to reconstruct the (partial) social network underlying the cohort. They then leveraged this social network data to study peer effects for obesity (Christakis and Fowler, 2007), smoking (Christakis and Fowler, 2008), and happiness (Fowler and Christakis, 2008). Researchers have since used the same methods as C&F to study peer effects in the FHS and in many other social networks settings. However, like epidemiologists studying cardiovascular disease, C&F and other researchers using non-experimental data to assess peer effects generally use statistical models that assume independence across subjects (Lyons, 2011); e.g. Trogdon et al. (2008); Fowler and Christakis (2008); Rosenquist et al. (2010). To assess peer influence for obesity, C&F fit longitudinal logistic regression models of each individual's obesity status at exam $k = 2, 3, 4, 5, 6, 7$ onto each of the individual's social contacts' obesity statuses at exam k and $k - 1$ (with a separate entry into the model for each contact), controlling for individual covariates and for the node's own obesity status at exam $k - 1$. They used generalized estimating equations (Liang and Zeger, 1986) to account for correlation within individual, but their model assumes independence across individuals. Christakis and Fowler fit this model separately for ten different types of social connections, including siblings, spouses, and immediate neighbors.

We replicated a secondary analysis in which the social contacts' obesity statuses at exams $k - 1$ and $k - 2$ were used instead of k and $k - 1$; we replicated

this analysis to avoid the misspecification inherent in the former specification (Lyons, 2011). Although it would be possible to adapt our proposed test of dependence to longitudinal or clustered data, that is beyond the scope of this paper and for simplicity we fit the C&F model at a single time point and selected one social contact for each node in order to have one residual per individual. We chose to use exam 3 for the outcome data because it gave us the largest sample size. We looked at sibling relationships because this gives the largest number of ties in the underlying network compared to the other nine types of relationships considered by Christakis and Fowler and because we had a prior hypothesis that close genetic relationships would evince dependence in obesity status.

We calculated Moran's I for the outcome (obesity status in exam 3), the predictor of interest (sibling's obesity status in exam 2), and the residuals from the logistic regression of each node's exam 3 obesity status onto the node's own obesity status in exam 2, the sibling's obesity status in exam 2, the sibling's obesity status at exam 1, and covariates age, sex, and education. For the outcome $I_{std} = 7.10$ ($p < 0.01$) and for the exposure $I_{std} = 15.91$ ($p < 0.01$) (because BMI is a binary variable I is equivalent to Φ), suggesting that confounding by network structure could contribute to any apparent association between the outcome and the exposure of interest. $I_{std} = 2.76$ ($p < 0.01$) for the regression residuals, providing strong evidence that the i.i.d. assumption on which these

analyses rests may be invalid. Details of our analysis can be found in the Supplementary Information.

3.4 Discussion

As researchers across many scientific disciplines grapple with replication crises, many sources of artificially small p-values and inflated false positive rates have received attention, but the possible impact of network dependence has been overlooked. In this paper, we used simple tests for independence among observations sampled from a single network to demonstrate that many types of analyses using FHS data may have reported biased point estimates and artificially small p-values, standard errors, and confidence intervals due to unacknowledged network dependence.

A limitation to the widespread application of these tests is their reliance on social network information, which is not available in most studies that are not explicitly about networks. Except in pathological cases, missing data on network ties will affect the power but not validity of these tests, so adding information on one or two ties per subject to a data collection protocol would enable researchers to test for network dependence. When some of the network ties are familial, and when genetic data is available, as is the case in the FHS, techniques developed to control for confounding due to cryptic relatedness (Sil-

lanpää, 2011) may be helpful for estimating the unknown familial network structure and for controlling for confounding due to that structure.

Future work is needed on methods to account for dependence if it is detected. Although many statistical methods exist for dealing with dependent data, most of these methods are intended for spatial or temporal data, or, more broadly, for observations with positions in $\mathbb{R}^k$ and dependence that is related to Euclidean distance between pairs of points. The topology of a network is very different from that of Euclidean space, and careful work is needed to justify the use of existing methods for social network data and to develop new methods.

We recommend that researchers designing new studies with human subjects avoid recruiting from one or a small number of underlying social networks whenever possible, and researchers working with existing data should be aware of the possibility that social network dependence may undermine the use of i.i.d. models.

3.5 Appendix : Analysis of the Framingham Heart Study data

The publicly available data are divided into datasets for individuals with and without non-profit use (NPU) consent, and for each replication we selected the dataset with more eligible individuals or more observed network ties, ex-

cept for the peer influence analysis, where we merged both consent groups.

3.5.1 Confounding by network structure

We used four random outcomes (systolic blood pressure, employed or not, visited a doctor due to illness, diagnosis of corneal arcus) from the Offspring Cohort at Exam 5 with NPU consent. The number of non-missing observations and the number of edges are shown in Table 3.4.

	Systolic blood pressure	Employed	Visited doctor	Corneal arcus
Sample size (n)	1031	1021	1028	1019
The number of edges (m)	683	670	681	674

Table 3.4: The number of observations and of undirected edges sampled from the Offspring Cohort at Exam 5.

For each of these four outcomes, we used the $n \times n$ outcome-specific adjacency matrix $\mathbf{A}$ to simulate continuous network dependent covariates $(X_1, X_2, \ldots, X_n)$ conditional on $\mathbf{A}$ as follows:

$$(X_1, X_2, \ldots, X_n) \sim \mathbf{MVN}\left(\mu = (\mu_1, \mu_2, \ldots, \mu_n), \mathbf{\Sigma}_n\right), \tag{3.2}$$

where $\mu_i = 1$ if $\sum_{j=1}^{n} A_{ij} > 0$ and $\mu_i = -1$ otherwise ($i = 1, 2, \ldots, n$). A variance-covariance matrix $\mathbf{\Sigma}_n = [\sigma_{ij}]$ has a diagonal of 0.5, $\sigma_{ij} = \sigma_{ji} = 0.2$ if $A_{ij} = A_{ji} = 1$, and $\sigma_{ij} = \sigma_{ji} = 0.1$ otherwise ($i, j = 1, 2, \ldots, n;\ i \neq j$).

3.5.2 Cardiovascular disease epidemiology

Lauer et al. (Lauer et al., 1991) used data from individuals with echocardiograms between 1979 and 1983, which coincides with the period of the original cohort Exam 16 (1979 - 1982) and the offspring cohort Exam 2 (1979 - 1983). Because we require information on network ties in order to test for dependence, we will consider a subset of the data used in Lauer et al. (1991), namely the observations from the Original Cohort Exam 16 (1979 - 1982) and the Offspring Cohort Exam 2 (1979 - 1983) without NPU consent. Because the analysis in Lauer et al. (1991) is stratified by sex, we constructed sex-specific adjacency matrices using the network ties which were in existence at the start of the cohorts (1979) or were initiated no later than the end of the original cohort Exam 16 (1982), so that any network ties present between 1979 to 1982 are taken account in the adjacency matrices. Figure 3.2 describes the eligibility criteria we used.

Tables 3.5 through 3.8 give summary measures for the variables used in our analyses and report the coefficients from the models that we fit. These can be compared to the published summaries and models in the original papers; we concluded that the summaries and models are sufficiently similar to deem our replications successful.

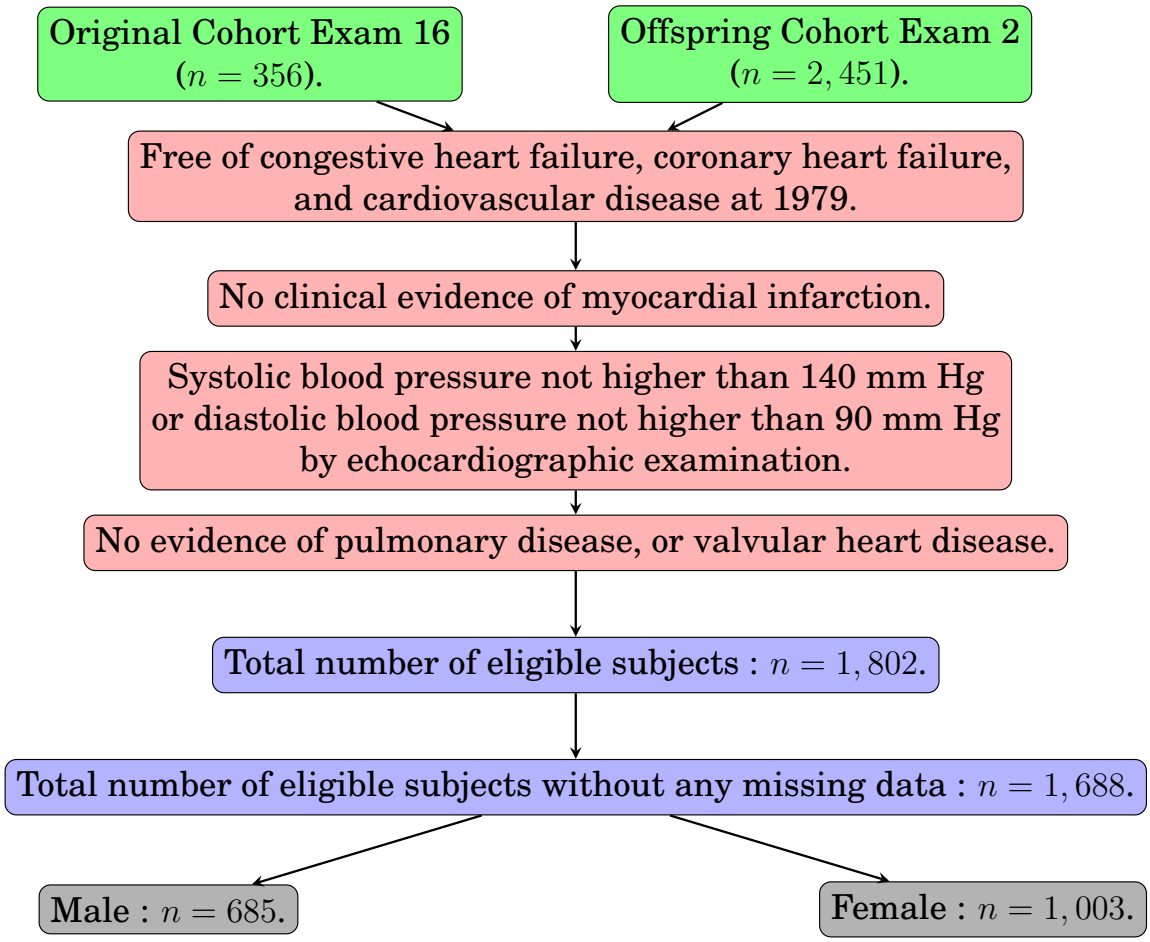
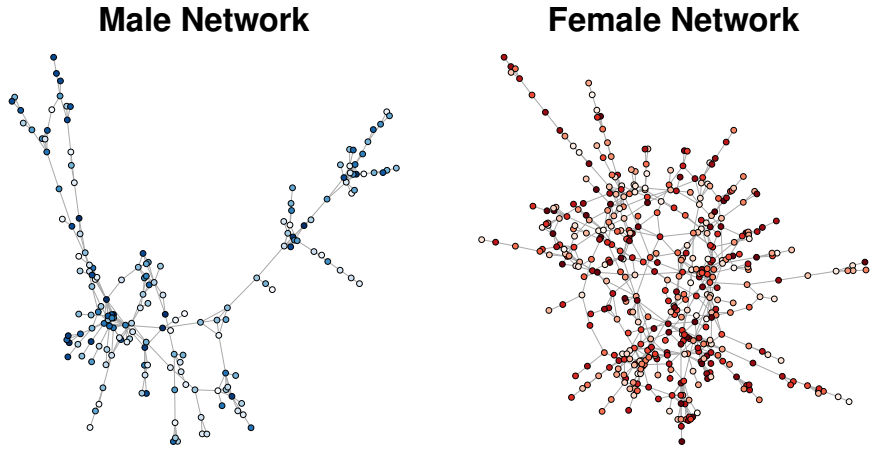


Figure 3.2: Flowchart for determining eligible healthy subjects from the Original and Offspring Cohorts consent group to replicate the left ventricular mass analysis of Lauer et al. (1991)

Table 3.5: Mean and standard deviations in the parenthesis of characteristics for eligible subjects. This corresponds to Table 1 in the original paper (Lauer et al., 1991) of left ventricular mass study.

	Male (n = 685)	Female (n = 1,003)
Age (year)	40.23 (10.04)	42.04 (10.67)
Weight (kg)	80.61 (10.86)	62.29 (11.10)
Height (cm)	177.2 (7.25)	162.3 (6.32)
BMI (kg/m^2)	2.33 (0.90)	1.75 (0.95)
Systolic BP ($mmHg$)	118.28 (9.41)	112.22 (11.20)
LVM (g)	207.27 (46.45)	135.2 (30.65)
Adjusted LVM (g/m)	116.85 (25.18)	83.3 (18.61)



(a) $I_{std} = 1.34$ (p-value : 0.090) (b) $I_{std} = 2.92$ (p-value : 0.002)

Figure 3.3: The largest connected components of the sex-specific social networks from left ventricular mass (LVM) study (Lauer et al., 1991), displayed using `Fruchterman-Reingold` algorithm. Darker colored nodes represent subjects with higher values of residuals from the regression of normalized LVM onto BMI, age, and systolic blood pressure.

Table 3.6: Replication of Lauer et al.'s linear regression of height-corrected left ventricular mass on BMI, age and systolic blood pressure.

	Estimate	Standard error	t-value	Pr(>|t|)
Male ($n = 685$)				
Intercept	70.46	11.35	6.21	0.00
Age	-0.16	0.09	-1.69	0.09
BMI : 23-25.99 kg/m^2	8.87	2.50	3.55	0.00
BMI : 26-29.99	20.43	2.57	7.95	0.00
BMI : ≥ 30	27.87	3.57	7.81	0.00
Systolic BP	0.34	0.10	3.40	0.00
Female ($n = 1,003$)				
Intercept	49.19	5.18	9.50	0.00
Age	0.20	0.05	3.89	0.00
BMI : 23-25.99	6.95	1.22	5.70	0.00
BMI : 26-29.99	15.02	1.61	9.34	0.00
BMI ≥ 30	27.97	2.02	13.86	0.00
Systolic BP	0.17	0.05	3.45	0.00

Table 3.7: Standard deviations of eight different heart rate variability measures from [Table 4] of the original paper (Tsuji et al., 1994) and from the 516 subjects we used to replicate the original analysis.

	ln**SDNN**	ln**pNN50**	ln**r-MSSD**	ln**VLF**	ln**LF**	ln**HF**	ln**TP**	ln**LF/HF**
Original paper	0.33	1.32	0.44	0.76	0.82	0.85	0.73	0.57
Our data	0.33	1.36	0.46	0.74	0.84	0.88	0.73	0.57

Table 3.8: Replication of twenty-four Cox models from [Table 5] in Tsuji et al. (Tsuji et al., 1994).

	Hazard ratio	95% CI	P-value
Unadjusted ($n = 516$)			
ln**SDNN**	1.31	(1.04, 1.64)	0.0217
ln**pNN50**	1.03	(0.81, 1.31)	0.8229
ln**r-MSSD**	1.05	(0.82, 1.34)	0.7092
ln**VLF**	1.53	(1.23, 1.90)	0.0001
ln**LF**	1.57	(1.25, 1.98)	0.0001
ln**HF**	1.27	(0.99, 1.64)	0.0607
ln**TP**	1.49	(1.20, 1.86)	0.0004
ln**LF/HF**	1.35	(1.08, 1.68)	0.0095
Age- and sex-adjusted ($n = 516$)			
ln**SDNN**	1.32	(1.06, 1.65)	0.0146
ln**pNN50**	1.14	(0.90, 1.45)	0.2781
ln**r-MSSD**	1.19	(0.94, 1.51)	0.1493
ln**VLF**	1.53	(1.23, 1.91)	0.0001
ln**LF**	1.56	(1.24, 1.97)	0.0002
ln**HF**	1.35	(1.06, 1.72)	0.0150
ln**TP**	1.51	(1.21, 1.89)	0.0003
ln**LF/HF**	1.17	(0.93, 1.46)	0.1911
Age, sex, and clinical risk factors adjusted ($n = 512$)			
ln**SDNN**	1.29	(1.02, 1.62)	0.0312
ln**pNN50**	1.13	(0.88, 1.44)	0.3480
ln**r-MSSD**	1.20	(0.94, 1.54)	0.1425
ln**VLF**	1.55	(1.24, 1.95)	0.0002
ln**LF**	1.49	(1.16, 1.92)	0.0018
ln**HF**	1.31	(1.03, 1.68)	0.0304
ln**TP**	1.51	(1.19, 1.90)	0.0006
ln**LF/HF**	1.10	(0.86, 1.41)	0.4570

Tables 3.9 through 3.11 summarize the results of tests for network dependence applied to the outcomes, primary covariates of interest, and residuals from the models of the three studies (Wolf et al., 1991; Gordon et al., 1977; Levy et al., 1990) that we did not include in the main text.

Table 3.9: Wolf et al. (Wolf et al., 1991) estimated the association between atrial fibrillation (AF) on sex- and age-group-specific two-year incidence of stroke, controlling for coronary heart disease, hypertension, and cardiac failure history. We replicated the analyses using data from the Original Cohort Exam 17 without NPU consent (the original study combined data from 17 exams). Below we report Moran's I statistic and the corresponding permutation-based p-values for the outcome (stroke), the predictor of interest (AF), and the residuals from the full logistic regression model.

	Stroke		**AF**		**Residuals**		n
	I_{std}	p-value	I_{std}	p-value	I_{std}	p-value	
Male							
60-69 yr	-0.19	0.460	0.22	0.152	-0.08	0.408	228
70-79 yr	-0.05	0.304	-0.10	0.438	0.01	0.274	267
80-89 yr	-0.85	0.908	-0.67	0.794	-0.95	0.950	93
Female							
60-69 yr	-1.02	0.984	-0.01	0.690	-0.67	0.942	258
70-79 yr	-0.12	0.544	0.04	0.334	0.15	0.368	398
80-89 yr	1.09	0.120	-0.06	0.410	-0.10	0.476	179

Table 3.10: Gordon et al. (1977) examined the association between HDL cholesterol and four-year incidence of coronary heart disease (CHD) for men and women aged 49 to 82 years old between 1969 and 1971, which coincides with Original Cohort Exam 11. We used the Original Cohort Exam 11 (with NPU consent group) to replicate their univariate logistic regressions of CHD on HDL, and below we report the network dependence test statistics and corresponding permutation-based p-values for the outcome, the predictor, and the residuals. (Due to the large amount of missingness in HDL, the statistics for the residuals are based on smaller sample sizes.)

Y	Sex	n	Moran's I_{std}	P-value
Four-year incidence of CHD	Male	1123	0.32	0.350
	Female	1416	-0.60	0.704
High density lipoproteins (HDL)	Male	552	1.64	0.042
	Female	640	2.05	0.030
Residuals from logistic regression	Male	552	-1.10	0.952
	Female	640	-0.10	0.524

Table 3.11: Levy at al. (Levy et al., 1990) investigated the relationship between left ventricular mass (LVM) and cardiovascular disease (CVD) for subjects 40 years old or older. We replicated their analyses of four-year incidence of CVD by running the logistic regression adjusted for age, diastolic blood pressure, pulse pressure, antihypertensive treatment, the number of cigarettes per day, diabetes status, body-mass index, ratio of total to high-density lipoprotein cholesterol, left ventricular hypertrophy on echocardiography, and left ventricular mass; below we report tests of network dependence and corresponding permutation-based p-values for the outcome (CVD), the predictor of interest (LVM), and the regression residuals. As in the original study we used observations ($n = 469$ males and $n = 713$ females) from the Original Cohort Exam 16 and the Offspring Cohort Exam 12, but we restricted our sample to those with NPU consent only.

Sex	Y	Moran's I_{std}	P-value
Male	Incidence of CVD	-0.63	0.744
Female	Incidence of CVD	0.74	0.210
Male	LVM	1.87	0.046
Female	LVM	1.21	0.146
Male	Residuals	-1.10	0.912
Female	Residuals	-0.20	0.450

Chapter 4

Network Dependence Testing via Diffusion Maps and Distance-Based Correlations

Deciphering the associations between network connectivity and nodal attributes is one of the core problems in network science. The dependency structure and high-dimensionality of networks pose unique challenges to traditional dependency tests in terms of theoretical guarantees and empirical performance. We propose an approach to test network dependence via diffusion maps and distance-based correlations. We prove that the new method yields a consistent test statistic under mild distributional assumptions on the graph structure, and demonstrate that it is able to efficiently identify the most informative

graph embedding with respect to the diffusion time. The testing performance is illustrated on both simulated and real data.

This is a joint work in collaboration with Chencheng Shen, Carey E. Priebe, and Joshua T. Vogelstein.

4.1 Introduction

Network data has seen increased availability and influence which motivated numerous recent advances of statistics and applications in physics, computer science, biology, social science, and more. However, data scientists and statisticians still confront many new and exciting challenges due to the distinct structure of network data. A network (or graph) is formally defined as an ordered pair $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ represents the set of nodes (or vertices) and $\mathbf{E}$ is the set of edges (or links), and $n = |\mathbf{V}|$. The edge connectivity of a graph can be compactly represented by the adjacency matrix $\mathbf{A} = \{\mathbf{A}(i, j) : i, j = 1, .., n\}$, where $\mathbf{A}(i, j)$ is the edge weight between node i and node j, e.g., for an unweighted and undirected network, $\mathbf{A}(i, j) = \mathbf{A}(j, i) = 1$ if and only if node i and node j are connected by an edge, and zero otherwise. In addition to edges, each node has a nodal attribute, denoted $X_i \in \mathbb{R}^p$, and $\mathbf{X} = [X_1 | \cdots | X_n]$.

This chapter focuses on testing the relationship between network connectivity and nodal attributes. Each node often has an associated nodal attribute,

and we would like to test whether the attributes are independent of the graph topology. Assuming for the adjacency matrix A and attributes X, the connectivity and attribute corresponding to each vertex are identically and jointly distributed according to F_{AX}, the null and alternative hypotheses of interest are:

$$H_0 : F_{AX} = F_A F_X \tag{4.1}$$
$$H_A : F_{AX} \neq F_A F_X.$$

This independence test is a crucial first step in exploring many network data, e.g. determining potential correlation between cultural tastes and relationships over social network (Lewis et al., 2012), or identifying association between the strength of functional connectivity and brain physiology (e.g., regional cerebral blood flow) in brain network (Liang et al., 2013). Sometimes the correlations among nodes are not proportional to the strength of connectivity between them. For instance, in signaling network of biological cells, reaction rate for each cell exhibits nonlinear dependence on the neighboring response due to complex, cooperative biological process involved (Hernandez-Hernandez et al., 2017). As an another example, in social network analysis, rumors may propagate at rates dependent on a few focal persons, rather than strength of connectivity (Nekovee et al., 2007).

A notable obstacle in network inference is the structure of the edge connectivity, e.g., for an undirected graph, A is a symmetric binary matrix where edges are not independent of each other, which prevents many well-established methods from being directly applicable. One approach is to assume certain model on the graph structure, and then solve the inference question based on the model assumption (Wasserman and Pattison, 1996; Fosdick and Hoff, 2015; Howard et al., 2016). Another approach is spectral embedding, which first embeds the $n \times n$ adjacency matrix A into an $n \times q$ matrix U by eigendecomposition, and then carries out the later inference task on U (Rohe et al., 2011; Sussman et al., 2012; Tang et al., 2017). For example, the network dependence test proposed by Fosdick and Hoff (Fosdick and Hoff, 2015) assumes that the adjacency matrix is generated from a multivariate normal distribution of the latent factors, and estimates the latent factor associated with each node from A, followed by applying the standard likelihood ratio test on the normal distribution.

However, model-based approaches are often limited by, and do not perform well beyond, the model assumptions. Moreover, spectral embedding is susceptible to misspecification of the dimension of q. Both of these factors can significantly degrade the later inference performance. Indeed, as a ground truth is unlikely in real networks (Peel et al., 2017), one often desires a method that is effectively non-parametric and robust against algorithm parameter selection (Chen et al., 2016).

We propose a method to test network dependency via diffusion maps and distance (or kernel)-based correlations, which is theoretically consistent under mild graph distributional assumptions (see Section 4.3.2), and works well for many popular network models. The proposed method also overcomes parameter selection issues, and exhibits superior empirical testing performance. The `R` code and accompanying data are publicly available online at `http://neurodata.io/tools/mgc` and `https://github.com/neurodata/mgc`.

4.2 Preliminaries

4.2.1 Notation

We denote a random variable by capital letter X with distribution F_X. For each node $i \in \mathbf{V}$, its attribute is denoted by X_i whose realizations are in $\mathbb{R}^p$, and its edge connectivity vector is denoted by A_i, with realizations in $\mathbb{R}^n$ constructing $n \times n$ adjacency matrix $\mathbf{A}$. We assume that $(X_i, A_i) \sim F_{XA}$, i.e., identically distributed attributes and connectivity vectors. Later we introduce a multiscale node-wise representation of the nodes as an $n \times q$ matrix $\mathbf{U}^t = [U_1^t | U_2^t | \cdots | U_n^t]$ for any $t \in \{0\} \bigcup \mathbb{Z}^+$, where q is the embedding dimension and t is the Markov iteration time step. Let $\cdot^*$ denote estimated optimality; $\cdot^t$ denotes either the t^{th} power or time step, which shall be clear in the context;

and $\cdot^T$ is the matrix transpose.

4.2.2 Diffusion maps

Because the rows and columns of a symmetric adjacency matrix may be correlated, directly operating on the adjacency matrix breaks theoretical guarantees of existing dependence tests. The diffusion map is introduced as a feature extraction algorithm by Coifman and Lafon (Coifman et al., 2005; Coifman and Lafon, 2006; Lafon and Lee, 2006), which computes a family of embeddings in Euclidean space by eigendecomposition on a diffusion operator of the given data. Here we introduce a version tailored to adjacency matrices.

To derive the diffusion maps for given observations of size n, the first step is to choose a $n \times n$ kernel matrix $\mathbf{K}$ that represents the similarity within the sample data. The adjacency matrix $\mathbf{A}$ is a natural similarity matrix; for undirected graphs we let $\mathbf{K} = \mathbf{A}$, for directed graphs we let $\mathbf{K} = (\mathbf{A} + \mathbf{A}^T)/2$. Next compute the normalized Laplacian matrix by

$$\mathbf{L} = \mathbf{B}^{-1/2}\mathbf{K}\mathbf{B}^{-1/2}, \tag{4.2}$$

where $\mathbf{B}$ is the $n \times n$ degree matrix of $\mathbf{K}$. When $\mathbf{B}(i,i)$ or $\mathbf{B}(j,j)$ is zero, $\mathbf{L}(i,j) = 0$.

The diffusion map $\mathbf{U}^t = \{U_i^t \in \mathbb{R}^q : i = 1, \ldots, n\}$ is then computed by eigen-

decomposition, namely

$$U_i^t = \left(\lambda_1^t \phi_{i1}, \quad \lambda_2^t \phi_{i2}, \quad \cdots, \quad \lambda_q^t \phi_{iq}\right)^T \in \mathbb{R}^q; \quad i = 1, \ldots, n, \tag{4.3}$$

where $\{\lambda_j^t : j = 1, 2, \ldots, q\}$ and $\{\phi_j \in \mathbb{R}^n : (\phi_{1j}, \phi_{2j}, \ldots, \phi_{nj}),\ j = 1, 2, \ldots, q\}$ are the q largest eigenvalues and corresponding eigenvectors of $\mathbf{L}$ respectively, and λ_j^t is the t^{th} power of the j^{th} eigenvalue. The diffusion distance between the i^{th} observation and the j^{th} observation is defined as the weighted ℓ^2 distance of the two points in the observation space, which equals the Euclidean distance in the diffusion coordinate:

$$\mathbf{C}^t(i, j) = \|U_i^t - U_j^t\|; \quad i, j = 1, 2, \ldots, n, \tag{4.4}$$

where $\|\cdot\|$ is the Euclidean distance.

When $t = 0$, the diffusion map is exactly the same as a normalized graph Laplacian embedding in Rohe et al. (2011) up-to a linear transformation; when $t > 0$, the diffusion maps are weighted graph Laplacian by powered eigenvalues (Lafon and Lee, 2006); and the diffusion map at $t = 1$ equals the adjacency spectral embedding (ASE) up-to the degree constant (Sussman et al., 2014). Therefore, the diffusion maps can be viewed as a single index family of embeddings. The embedding dimension choice q can be selected via the profile likelihood method in Zhu and Ghodsi (2006), which is a standard algorithm

in dimension reduction literature. To select the optimal t, we will utilize a smoothing technique to maximize the dependency, as discussed below.

4.2.3 Distance-based correlations

The problem of testing general dependencies between two random variables has seen notable progress in recent years. The Pearson's correlation (Pearson, 1895) is the most classical approach, which determines the existence of linear relationship via a correlation coefficient in the range of $[-1, 1]$, with 0 indicating no linear association and ± 1 indicating perfect linear association. To better capture the dependencies not limited to linear relationship, a variety of distance-based correlation measures have been suggested, including the Mantel coefficient (Mantel, 1967), distance correlation (DCORR) and energy statistic (Székely et al., 2007; Székely and Rizzo, 2013a; Rizzo and Székely, 2016), kernel-based independence test (Gretton and Gyorfi, 2010), Heller-Heller-Gorfine (HHG) test (Heller et al., 2013, 2016), and multiscale graph correlation (MGC) (Shen et al., 2018a,b), among others. In particular, distance correlation is a distance-based dependency measure that is consistent against all possible dependencies with finite first moments. The multiscale graph correlation (MGC) statistic inherits the same consistency of distance correlation with remarkably better finite-sample testing powers under high-dimensional and nonlinear dependencies, via defining a family of local

correlations and efficiently searching for the optimal local scale in testing. Here we briefly introduce DCORR and MGC.

Given n pairs of sample data $(\mathbf{U}, \mathbf{X}) = \{(U_i, X_i) \overset{i.i.d.}{\sim} F_{UX} \in \mathbb{R}^q \times \mathbb{R}^p : i = 1, 2, \ldots, n\}$. Denote the pairwise distances within $\{U_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ as $\mathbf{C}(i,j) = \|U_i - U_j\|$ and $\mathbf{D}(i,j) = \|X_i - X_j\|$ for $i, j = 1, 2, \ldots, n$ respectively. The sample distance covariance is defined as

$$\text{DCOV}_n(\mathbf{U}, \mathbf{X}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{\mathbf{C}}(i,j)\tilde{\mathbf{D}}(i,j), \tag{4.5}$$

where $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{D}}$ doubly-center $\mathbf{C}$ and $\mathbf{D}$ by their column means and row means, respectively, i.e., $\tilde{\mathbf{C}} = \mathbf{HCH}$, where $\mathbf{H} = \mathbf{I}_{n\times n} - \mathbf{J}_{n\times n}/n$ (the double centering operation matrix), $\mathbf{I}_{n\times n}$ is the $n \times n$ identity matrix (ones on the diagonal, zeros elsewhere), and $\mathbf{J}_{n\times n}$ is the $n \times n$ matrix of all ones. The distance correlation (DCORR) follows by normalizing distance covariance via Cauchy-Schwarz into the range of $[-1, 1]$ (i.e., divide by $\{\sum_{i,j=1}^{n} \tilde{\mathbf{C}}(i,j)^2/n^2 \sum_{i,j=1}^{n} \tilde{\mathbf{D}}(i,j)^2/n^2\}^{1/2}$). Székely et al. (2007) shows that sample DCORR converges to a population form, which is asymptotically 0 if and only if independence, i.e., $F_{UX} = F_U F_X$, resulting in a consistent statistic for independence testing; an unbiased sample version of distance correlation is later proposed to eliminate the sample bias in DCORR (Székely and Rizzo, 2013b, 2014), which is the default for DCORR implementation in this chapter.

The MGC statistic is an optimal local version of distance correlation, aiming at improving finite-sample testing power. It first derives all local distance covariances DCOV^{kl} as

$$\text{DCOV}_n^{kl}(\mathbf{U}, \mathbf{X}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{\mathbf{C}}^k(i,j)\tilde{\mathbf{D}}^l(i,j); \quad k = 1, \ldots, \kappa,\ l = 1, \ldots, \gamma, \tag{4.6}$$

where κ and γ are the number of unique numerical values in $\mathbf{C}$ and $\mathbf{D}$ respectively; $\tilde{\mathbf{C}}^k(i,j) = \tilde{\mathbf{C}}(i,j)\mathbf{I}(R_{ij}^{\mathbf{C}} \leq k)$; $\mathbf{I}(\cdot)$ is the indicator function; and $R_{ij}^{\mathbf{C}}$ is a rank function of U_i relative to U_j, i.e., $R_{ij}^{\mathbf{C}} = k$ if U_i is the k^{th} nearest neighbor of U_j, and define equivalently $\tilde{\mathbf{D}}^l(i,j) = \tilde{\mathbf{D}}(i,j)\mathbf{I}(R_{ij}^{\mathbf{D}} \leq l)$ for $\{X_i\}$. Then the local distance correlations $\{(\text{DCORR}^{kl})\}$ are the normalizations of the local distance covariances into $[-1, 1]$ via Cauchy-Schwarz. Among all possible neighborhood choices, the MGC statistic equals the maximum local correlation within the largest connected component of significant local correlations, i.e.,

$$\text{MGC}_n(\mathbf{U}, \mathbf{X}) = \text{DCORR}_n^{(kl)^*}(\mathbf{U}, \mathbf{X}), \text{ where } (kl)^* = \arg\max_{(kl)} \mathbf{S}(\text{DCORR}_n^{kl}) \tag{4.7}$$

for a smoothing operation $\mathbf{S}(\cdot)$ that filters out all in-significant local correlations.

MGC has been shown to have power almost equal or better than DCORR throughout various types of general dependencies. Despite searching over all possible neighborhoods for the optimal local correlation, it is also computationally effi-

cient with a similar running time complexity as DCORR and HHG. The details on the population statistic, sample version on unbiased DCORR, and running time analysis are described in Shen et al. (2018a).

4.3 Main Results

4.3.1 Testing procedure of diffusion MGC

Algorithm 1 Testing procedure of DMGC.

Input: Adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and nodal attributes $\mathbf{X} = \{X_i \in \mathbb{R}^p : i = 1, 2, \ldots, n\}$.
(1) Symmetrize $\mathbf{A}$ by $\mathbf{K} = (\mathbf{A} + \mathbf{A}^T)/2$.
(2) Obtain normalized graph Laplacian matrix $\mathbf{L} = \mathbf{B}^{-1/2}\mathbf{K}\mathbf{B}^{-1/2}$.
(3) Do eigendecomposition to obtain diffusion maps $\mathbf{U}^t = \{U_1^t, U_2^t, \ldots, U_n^t\}$ for $t = 0, 1, 2, \ldots, 10$.
(4) Derive $n \times n$ Euclidean distance of diffusion map $\mathbf{C}^t$, i.e., diffusion distance, across t, and $n \times n$ Euclidean distance of nodal attributes, $\mathbf{D}$.
(5) Compute MGC statistics using two distance matrices, $\mathbf{C}^t$ and $\mathbf{D}$, for $t = 0, 1, \ldots, 10$.
(6) Derive DMGC statistic $\text{MGC}_n^* (\{\mathbf{U}^t\}, \mathbf{X})$ by estimating t^*.
(7) Compute p-value using permutation test.

Output: P-value at the estimated optimal step t^*, the estimated optimal time step t^*, dimension choice of q via profile likelihood method, multiscale local correlation maps $\{\text{DCORR}_n^{kl}(\mathbf{U}^t, \mathbf{X})\}$, the optimal neighborhood choice (k^*, l^*).

Here we develop diffusion MGC (DMGC), which synthesizes diffusion map embedding as a node-wise representation and distance-based measure by MGC as a test statistic to detect the signal using smoothed maximum statistic. A flowchart of the testing procedure is illustrated in Figure 4.1. The algorithm

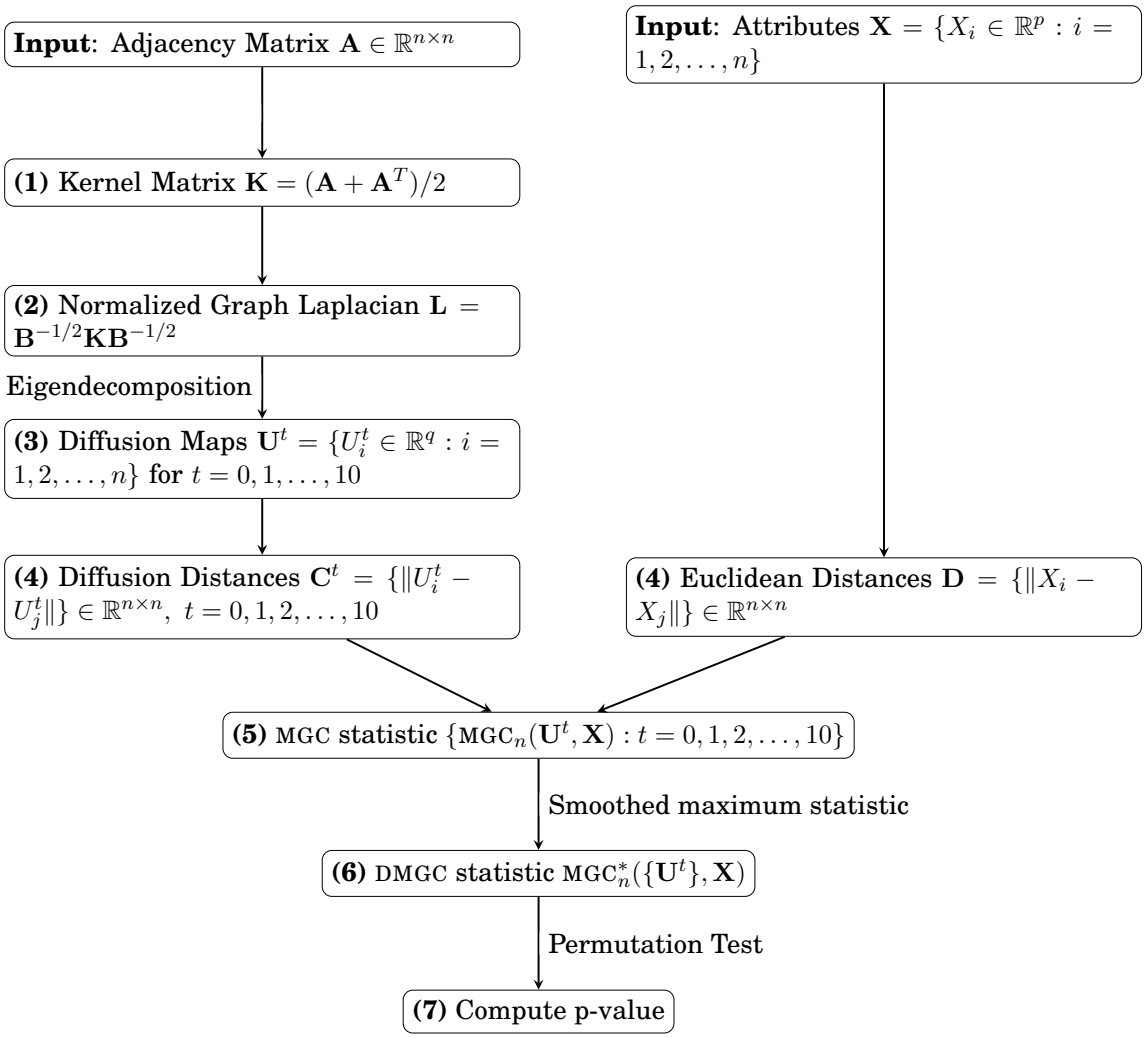


Figure 4.1: Flowchart for network dependence testing via diffusion maps and MGC (DMGC).

is flexible in the choice of correlation measures: by following the exact same steps except replacing MGC by DCORR, HHG, or another correlation measure in Step **(5)**, one can compute the diffusion DCORR or diffusion HHG statistic. The details for each step are described in Algorithm 1. In selecting t^* among multiscale statistics, the motivation of the smoothing in Step **(6)** is the following: suppose that edge connectivity is dependent with the attributes and

there exists an optimal t for detecting the relationship, then the test statistic at adjacent time steps should also exhibit strong signal, because the level of connectivity considering all paths of length t between each pair is most similar to those of length $t-1$ or $t+1$. Under independence, a large test statistic at certain t can occur by chance and cause a direct maximum to have a low testing power, while the smoothed maximum effectively filters out any noisy and isolated large test statistic. In practice, it suffices to consider $t \in [0, 1, \ldots, 10]$ or even smaller upper bound like 3 or 5. When smoothed maximum does not exist, we set $t = 3$ as the default choice.

The permutation test in Step **(7)** is a common nonparametric procedure used for real data testing in almost all dependency measures (Székely et al., 2007; Heller et al., 2013; Gretton and Gyorfi, 2010; Shen et al., 2018a), which is valid as long as the observations are exchangeable under the null (Rizzo and Székely, 2016). Because the null distribution of any correlation measure depends on the marginal distribution and is difficult to obtain, the permutation test breaks the dependence of the given data while keeping the marginal distributions, and is thus able to empirically approximate the null distribution.

4.3.2 Theoretical properties under exchangeable graph

To derive the theoretical consistency of our methodology, the following distributional assumptions on the distribution of the graph and the nodal attributes are required:

(C1) Graph $\mathbf{G}$ is an induced subgraph of an infinitely exchangeable graph, i.e., the adjacency matrix $\mathbf{A}$ satisfies

$$\mathbf{A}(i,j) \stackrel{d}{=} \mathbf{A}(\sigma(i),\sigma(j)) \tag{4.8}$$

for any $i,j = 1,2,\ldots,n$ and any permutation σ of size $n \in \mathbb{N}$. The notation $\stackrel{d}{=}$ stands for equality in distribution.

(C2) Each nodal attribute X_i is generated independently and identically from F_X with finite first moment.

(C3) The matrix $\mathbf{A}$ is constrained to a domain Ω, such that the diffusion map embedding from $\mathbf{A} \in \Omega$ to $\mathbf{U}^t$ is injective for some t.

Condition (C1) states that $\mathbf{G}$ is a collection of independently sampled nodes and their induced subgraph (Orbanz and Roy, 2015; Tang et al., 2017; Orbanz, 2017); this is a distributional assumption satisfied by many popular statistical networks models. Based on condition (C1), the diffusion map $\mathbf{U}^t$ at each t can

furnish exchangeable and asymptotic conditional i.i.d. embedding for the set of nodes $\mathbf{V}(\mathbf{G})$, under which the permutation test is valid.

Theorem 1. *Assume* $\mathbf{G}$ *satisfies (C1). Then at each fixed* t*, the diffusion maps* $\mathbf{U}^t = \{U_i^t,\ i = 1, 2, \ldots, n\}$ *embedded from the adjacency matrix* $\mathbf{A}$ *as Equation 4.3 are exchangeable. As a result, there exists an underlying variable* θ^t *distributed as the limiting empirical distribution of* $\mathbf{U}^t$*, such that* $U_i^t|\theta^t$ *are i.i.d. for* $i = 1, 2, \ldots, n$ *as* $n \to \infty$*.*

This exchangeability and asymptotically conditionally i.i.d. property of the diffusion map leads to the consistency of DMGC for testing independence between $\mathbf{U}^t$ and $\mathbf{X}$. Moreover, due to condition (C1), the permutation test is applicable to any $\mathbf{U}^t$ from an exchangeable sequence. In that sense, condition (C2) is merely a regularity condition; while the distribution of U_i^t always satisfies the finite-moment assumption (shown in the Supplementary Material for proof of Theorem 2). The condition (C1) and (C2) lead to the consistency of the intermediate test statistic between the diffusion map at each t and the nodal attribute.

Theorem 2. *Assume the graph* $\mathbf{G}$ *and the nodal attributes satisfy condition (C1) and (C2). Then as* $n \to \infty$*, the* MGC *statistic between the diffusion map* $\mathbf{U}^t$ *at any fixed* t *and the nodal attributes* $\mathbf{X}$ *satisfies:*

$$\text{MGC}_n(\mathbf{U}^t, \mathbf{X}) \to c \geq 0, \tag{4.9}$$

where equality holds if and only if $F_{U^tX} = F_{U^t}F_X$*, where each observation in* $\mathbf{U}^t$ *is identically distributed as* F_{U^t}.

The consistency of DMGC follows in the next theorem, which is extended to consistency between edge connectivity and nodal attributes if condition (C3) is satisfied. Condition (C3) connects the dependence test between $\mathbf{X}$ and $\mathbf{U}^t$ to the test between $\mathbf{X}$ and $\mathbf{A}$ in Equation 4.1.

Theorem 3. *Under the same assumption in Theorem 2, it holds that*

$$\text{MGC}_n^*(\{\mathbf{U}^t\}, \mathbf{X}) \to c \geq 0, \tag{4.10}$$

with equality holds if and only if $F_{U^tX} = F_{U^t}F_X$ *for all* $t \in [0, 10]$. *Therefore,* DMGC *is a valid and consistent statistic for testing independence between the diffusion maps* $\{\mathbf{U}^t\}$ *and nodal attributes* $\mathbf{X}$.

If condition (C3) holds, then $\text{MGC}_n^*(\{\mathbf{U}^t\}, \mathbf{X})$ *is also valid and consistent for testing independence between the adjacency matrix and nodal attributes, i.e., it converges to* 0 *if and only if the nodal attribute* X *is independent of the node connectivity* A.

Interpreted in another way, DMGC is always valid to use under Condition (C1) and (C2), but may not always detect the dependency between the edge connectivity and nodal attributes if the dependency signal is lost during the diffusion map embedding procedure. Therefore, condition (C3) on injective

transformation is a sufficient condition to preserve the dependency signal for diffusion maps.

Corollary 1. *Theorem 3 still holds, when any of the following changes are applied to the testing procedure described in Section 4.3.1:*

(1) The MGC *statistic in step 2 is replaced by sample* DCORR *or* HHG*;*

(2) When $\mathbf{A}$ *is restricted to be symmetric, binary, and of finite rank* $q < n$*, then condition (C3) holds at* $t = 1$*.*

Namely, point (1) suggests that under diffusion maps, consistent distance-based measures such as DCORR and HHG can also be used instead of MGC; this enables us to compare DMGC to diffusion DCORR and HHG in the simulations. And point (2) offers an example of random matrix $\mathbf{A}$ where the diffusion map is guaranteed injective within the domain.

4.3.3 Consistency under random dot product graph

In this section, we consider the random dot product graph model (RDPG), which is widely used in network modeling and ideal for illustration. A graph generated from RDPG model has an edge probability as a dot product of i.i.d. node-wise latent position: assuming each node has a latent position $W_i \overset{i.i.d.}{\sim} F_W$ for $i = 1, 2, \ldots, n$, the edge probability $pr(\mathbf{A}(i, j) = 1 \mid W_i, W_j)$ is determined by

the dot product of the latent positions, i.e.,

$$\mathbf{A}(i,j) \mid W_i, W_j \overset{i.i.d.}{\sim} Bern\big(\langle W_i, W_j \rangle\big),\ i,j = 1,2,\ldots,n \text{ and } i<j, \tag{4.11}$$

under the restriction that all W_i's are non-negative vectors and the dot product must be normalized within $[0,1]$.

An RDPG is an exchangeable graph model that satisfies condition (C1). In addition, RDPG fully specifies all exchangeable graph models that are unweighted and symmetric, whose probability generating matrix $\mathbf{P}(i,j) = \langle W_i, W_j \rangle$ is positive semi-definite.

Proposition 1 (Sussman et al. (2014))**.** *An exchangeable random graph has a finite rank q and positive semi-definite link matrix* $\mathbf{P}$*, if and only if the random graph is distributed according to a random dot product graph with i.i.d. latent vectors* $\{W_i \in \mathbb{R}^q, i = 1,\ldots,n\}$.

Indeed, many other popular network modelings are special cases of RDPG, including the stochastic block model, its degree-corrected version, the latent factor model from Fosdick and Hoff (2015), etc.

Proposition 2 (Rohe et al. (2011))**.** *Let* $\mathbf{L}$ *be the normalized graph Laplacian for an adjacency matrix* $\mathbf{A}$ *generated by an RDPG with latent positions of which construct the matrix of* $\mathbf{W} = [W_1|W_2|\ldots|W_n] \in \mathbb{R}^{q\times n}$*. Let* $\mathbf{U}^{t=1} = [U_1^{t=1}|U_2^{t=1}|\ldots|U_n^{t=1}] \in \mathbb{R}^{q\times n}$*. Then there exists a fixed diagonal matrix* $\mathbf{M}$ *and*

an orthonormal rotational matrix $\mathbf{Q} \in \mathbb{R}^{q \times q}$ *such that* $\|\mathbf{U}^{t=1} - \mathbf{QMW}\| \to 0$ *almost surely.*

Therefore, under RDPG, the diffusion map $\mathbf{U}^{t=1}$ asymptotically equals the latent position $\mathbf{W}$ up to a linear transformation. As the latent position under RDPG can be asymptotically recovered by diffusion maps, DMGC is consistent against testing general dependency between $\mathbf{A}$ and $\mathbf{X}$ under RDPG.

Corollary 2. *Under an induced subgraph from exchangeable graph with positive semi-definite link function, the* DMGC *statistic is consistent for testing independence between edge connectivity and nodal attributes.*

4.4 Numerical Studies

4.4.1 Stochastic block model

Throughout the numerical studies, we compare DMGC to the likelihood ratio test proposed by Fosdick and Hoff (FH) (Fosdick and Hoff, 2015), single embedding tests (using the adjacency spectral embedding (AM) and the latent factors (LF) with distance-based tests like (DCORR and HHG), as well as diffusion DCORR and diffusion HHG. The main approach of DMGC is denoted as MGC∘DM (or DMGC for brevity), and the benchmarks are FH, MGC/DCORR/HHG∘ AM/LF/DM.

For each simulation, we generate a sample graph and the corresponding attributes, compute the correlation measure on the respective embedding, carry out the permutation test with $r = 500$ random samples for each method, and reject the null if the resulting p-value is less than $\alpha = 0.05$. The testing power of each method equals the percentage of correct rejection out of $m = 100$ Monte-Carlo replicates, and a higher power implies a better method against the given dependency structure. The first simulation samples graphs from the stochastic block model (SBM) (Airoldi et al., 2008; Hanneke and Xing, 2009; Rohe et al., 2011; Xin et al., 2017). The SBM assumes that each of n nodes in $\mathbf{G}$ must belong to one of $K \in \mathbb{N}$ blocks, and determines the edge probability based on the block-membership of the connecting nodes: For $i = 1, \ldots, n$, assume that a latent variable of $Z_i \overset{i.i.d.}{\sim} Multinomial\big(\pi_1, \pi_2, ..., \pi_K\big)$ denotes the block-membership of each node, and $b_{kl} \in \{0, 1\}$ implies the edge probability between any two nodes of class k and l respectively; then the upper triangular entries of $\mathbf{A}$ are independently and identically distributed conditioned on $\mathbf{Z} = \{Z_i : \ i = 1, 2, \ldots, n\}$:

$$\mathbf{A}(i,j) \mid Z_i, Z_j \overset{i.i.d.}{\sim} Bernoulli\Big(\sum_{k,l=1}^{K} b_{kl}\mathbf{I}\big(Z_i = k, Z_j = l\big)\Big); \quad i < j, \ i, j = 1, 2, \ldots, n, \tag{4.12}$$

where $\mathbf{I}(\cdot)$ is the indicator function.

Our desire is to detect whether the adjacency structure is dependent on

nodal attributes, which here corresponds to the block assignment. Thus we consider testing dependency between the graph having the adjacency matrix $\mathbf{A}$ and a noisy block-membership $\mathbf{X}$, which are correlated through true block-membership $\mathbf{Z}$:

$$
\begin{aligned}
&Z_i \overset{i.i.d.}{\sim} Multinom(1/3, 1/3, 1/3),\\
&\mathbf{A}(i,j) \mid Z_i, Z_j \sim Bernoulli\left(0.5\mathbf{I}(|Z_i - Z_j| = 0) + 0.2\mathbf{I}(|Z_i - Z_j| = 1) + 0.4\mathbf{I}(|Z_i - Z_j| = 2)\right),\\
&X_i \mid Z_i \sim Multinom((1 + \mathbf{I}(Z_i = 1))/4,\ (1 + \mathbf{I}(Z_i = 2))/4,\ (1 + \mathbf{I}(Z_i = 3))/4),
\end{aligned}
\tag{4.13}
$$

and we set the sample size as $n = 100$. Equation 4.13 implies that the within-block edge probability is always 0.5; while the between-block edge probability is 0.2 when the block labels differ by 1, and 0.4 when the block labels differ by 2. A visualization of the sample data is shown in Figure 4.7(a). Note that nodal attributes $\mathbf{X}$ from the above model are the noisy version of the true block-membership by: for each i, $X_i = Z_i$ with probability 0.5, and equally likely to take other values in Ω, i.e., the true block-membership are observed half of the time. Notably, although within-block edge probability is the largest, the between-block edge probability is not linearly related to the distance of the block-membership, i.e., the edge probability between a node of block 1 and a node of block 3 is higher than the edge probability between block 1 and block 2. Therefore, this three-block SBM generates a noisy and nonlinear dependency

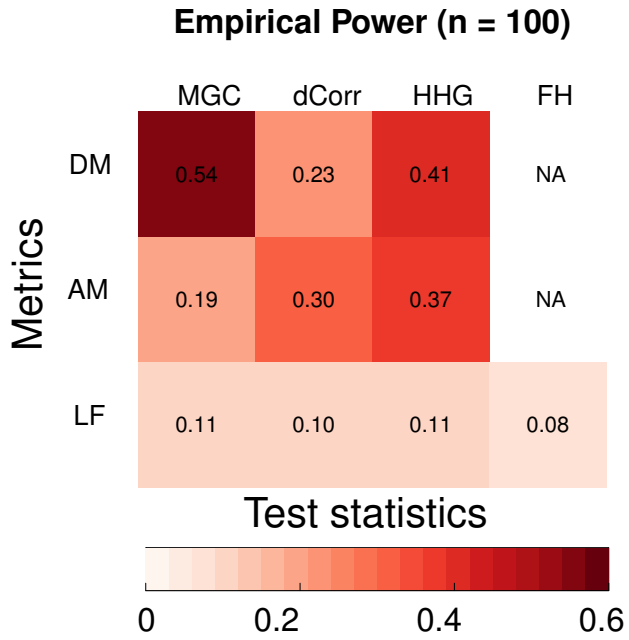


Figure 4.2: The power heatmap under the three-block SBM (Equation 4.13) demonstrates that among all possible combinations of test statistics with distance metrics, DMGC (top left entry) provides the best power.

structure between A and X. Here DMGC is expected to work better than all the other combinations mainly because MGC captures high-dimensional nonlinear dependencies better than DCORR, HHG, and the standard likelihood ratio test. Indeed, Figure 4.2 shows that DMGC prevails in the testing powers among all the methods.

4.4.2 SBM with linear and nonlinear dependencies

One of the many advantages in SBM is that we can easily create nonlinear dependency by manipulating a parameter. To better understand the advantage of our main approach (MGC $\circ$ DM) under different scenarios, here we use the same three-block SBM and its block-membership $\{Z_i : i = 1, 2, \ldots, n = 100\}$ as in the previous section, except that the edge probability is now controlled by $\beta \in (0, 1)$ as follows for all $i, j = 1, \ldots, n$:

$$\mathbf{A}(i,j) \mid Z_i, Z_j \sim Bernoulli\left(0.5\mathbf{I}(|Z_i - Z_j| = 0) + 0.2\mathbf{I}(|Z_i - Z_j| = 1) + \beta\mathbf{I}(|Z_i - Z_j| = 2)\right). \tag{4.14}$$

The noisy block-membership $\mathbf{X}$ is generated in the same way as before. When $\beta = 0.2$, the three-block SBM is the same as a two-block SBM, where within-block edge probability equals 0.5 while the between-block edge probability is always 0.2, i.e., it represents a linear association between the adjacency matrix and the block-membership; when $\beta < 0.2$, the association is still monotonic; when $\beta > 0.2$ and gets further away, the relationship becomes strongly nonlinear. Figure 4.3 plots the power against β for all diffusion maps-based methods. All of diffusion MGC, diffusion DCORR, and diffusion HHG perform almost the

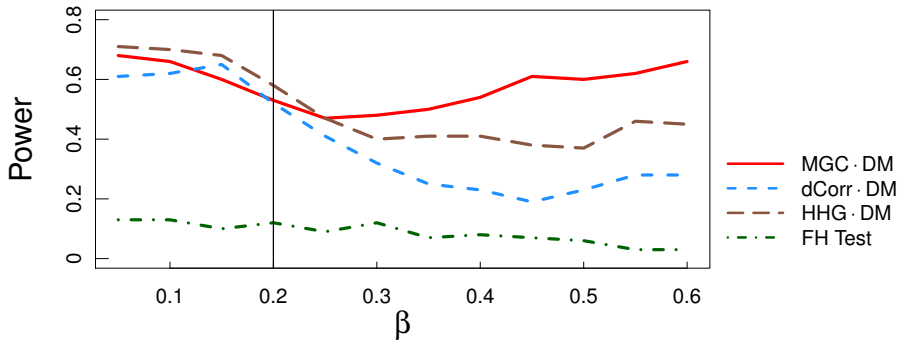


Figure 4.3: The power curve with respect to increasing β under three-block SBM (Equation 4.14). When $\beta \leq 0.2$, the dependency between the adjacency matrix and the block structure is linear or monotonic; when $\beta > 0.2$, the dependency becomes more and more nonlinear and non-monotonic as β gets further away from 0.2. Among all methods utilizing diffusion maps, MGC is evidently the best performing one when $\beta \geq 0.25$, implying that it better captures nonlinear dependencies. FH does not perform well for any value of β.

same at linear dependency (i.e, $\beta \leq 0.2$), with diffusion MGC (DMGC) being significantly more powerful as the dependency shifts to strong nonlinearity and even increasing as β increases. This observation demonstrates empirically that MGC better captures nonlinear dependencies in network testing for these settings.

4.4.3 Degree-corrected SBM

In this section we compare different embeddings under the degree-corrected stochastic block model (DC-SBM), which better reflects many real-world networks (Karrer and Newman, 2011). The DC-SBM is an extension of SBM by introducing an additional random variable c_i to control the degree of each node.

We set $n = 200$ with two blocks, select the binary block-membership Z_i uniformly in $\Omega = \{0, 1\}$, and generate the edge probability by

$$\mathbf{A}(i,j) \mid Z_i, Z_j, C_i, C_j \sim Bernoulli\left(0.2C_iC_j \cdot \mathbf{I}(|Z_i - Z_j| = 0) + 0.05C_iC_j \cdot \mathbf{I}(|Z_i - Z_j| = 1)\right), \tag{4.15}$$

where $C_i \overset{i.i.d.}{\sim} Uniform(1-\tau, 1+\tau)$ for $i = 1, \ldots, n$, and $\tau \in [0, 1]$ is a parameter to control the amount of variability in the edge degree, e.g., as τ increases, the model becomes more complex as the variability of the edge probability becomes larger; when $\tau = 0$, the above model reduces to a two-block SBM without any variability induced by $\{C_i : i = 1, 2, \ldots, n\}$.

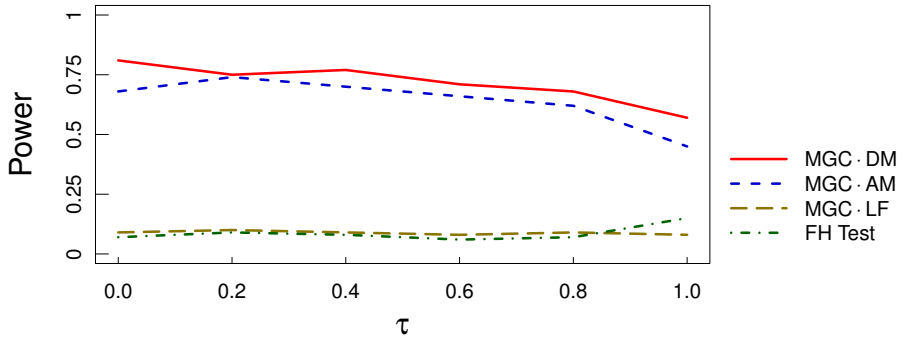


Figure 4.4: The power curve with respect to increasing τ under DC-SBM (Equation 5.3). The edge variability increases as τ does, the testing power of diffusion maps is relatively stable against increasing variability; the adjacency spectral embedding is slightly worse, while the latent positions fail to detect the dependency across all levels of τ.

We again generate the nodal attributes $\mathbf{X}$ as a noisy version of the true block-membership via Bernoulli distribution, i.e., for each i, $X_i = Z_i$ with prob-

ability 0.6, and equals the wrong label with probability 0.4. Figure 4.4 compares different embeddings with MGC, which shows that MGC$\circ$ DM (DMGC) and MGC$\circ$ AM have better testing performance over the other embedding methods for all values of τ.

4.4.4 RDPG simulations

Here we present a variety of RDPG simulations by generating the latent variables via the 20 relationships in Shen et al. (2018a) with different levels of noise, consisting of various linear, monotonic and non-monotonic (and therefore nonlinear) relationships. The details of simulation schemes are in the Supplementary Material while a general outline for data generating process is:

$$\begin{aligned} \begin{pmatrix} \tilde{W}_i & \tilde{X}_i \end{pmatrix} &\overset{i.i.d.}{\sim} F_{\tilde{W}\tilde{X}} \qquad i = 1, 2, \ldots, n, \\ \mathbf{A}(i,j) \mid W_i, W_j &\sim Bernoulli\left(\langle W_i, W_j \rangle\right), \; i < j = 1, 2, \ldots, n, \end{aligned} \tag{4.16}$$

where $W_i = (\tilde{W}_i - \min(\{\tilde{W}_j : j = 1, 2, \ldots, n\}))/(\max(\{\tilde{W}_j : j = 1, 2, \ldots, n\}) - \min(\{\tilde{W}_j : j = 1, 2, \ldots, n\})$ for $i = 1, 2, \ldots, n$, so that all the latent variable range from 0 to 1. We apply the same scaling from $\tilde{X}_i$ to X_i for visual consistency. Thus the latent positions and nodal attributes are correlated via a joint distribution of $F_{\tilde{W}\tilde{X}}$, which includes linear, quadratic, circle, and more. Figure 4.5 shows empirical power obtained from $m = 100$ independent replicates when

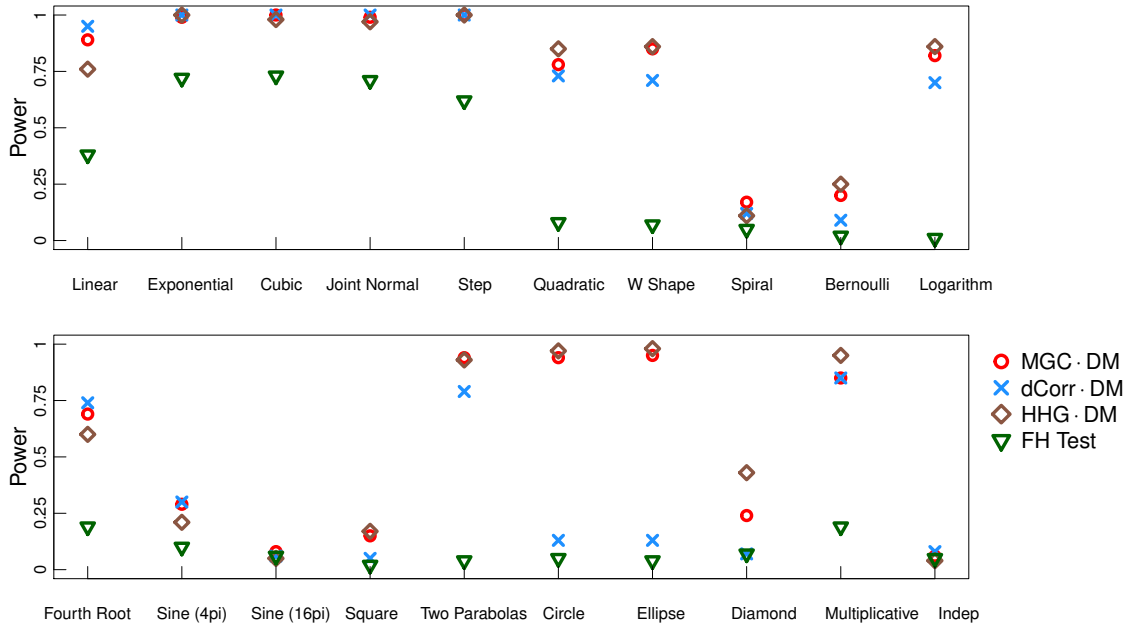


Figure 4.5: Power comparison for 20 different RDPGs with $n = 50$ nodes per $m = 100$ independent replicates. It shows that when latent positions W_i and nodal attributes X_i are dependent via a close-to-linear relationship (upper panel), all the distance-based tests achieve similar power while FH test is slightly worse due to its model-based nature. When non-linearity between W_i and X_i becomes evident like circle or ellipse (lower panel), DCORR and FH tests are far behind than MGC and HHG.

the number of nodes is $n = 50$. A lack of power in the FH test is evident even though data generative model in Equation 4.16 agrees with Fosdick and Hoff (2015)'s.

All the distance-based methods work fairly well, with diffusion MGC and diffusion HHG being the best performers. Note that the last scenario is an independent relationship and all tests achieve a power approximately at 0.05, implying that they are all valid tests; there are also a few dependencies of very low power due to the complexity of the relationship (sine, spiral, square, etc.),

but their powers all converge to 1 as sample size n increases.

4.5 DMGC Graph Embedding

This section demonstrates that in deriving DMGC, we preserve dependency structure between A and X without cross-validation or over-fitting by virtue of effectively estimating parameters of t and q.

As a reminder, the dimension choice q is selected by the second elbow of the absolute eigenvalue scree plot by the profile likelihood method from Zhu and Ghodsi (2006), which is a widely-used automatic algorithm for selecting the number of important features whenever eigenvalues or singular values are involved. The choice of t^* is based on a smoothed maximum, i.e., we take the maximum correlation only when consecutive MGC statistics are also large. Viewed in another way, DMGC selects the optimal diffusion map that maximizes the MGC statistic. Thus any testing advantage shall come down to whether it is able to optimize the embedding without over-fitting, and we investigate how well our procedure is able to preserve the dependency compared to using a single embedding choice.

Figure 4.6 presents the diffusion distances at different t and q for the three-block stochastic block model in Equation 4.13. Although the resulting embedding is sensitive to both t and q in Figure 4.6 (a)–(d), at optimal $t^* = 2$ it is

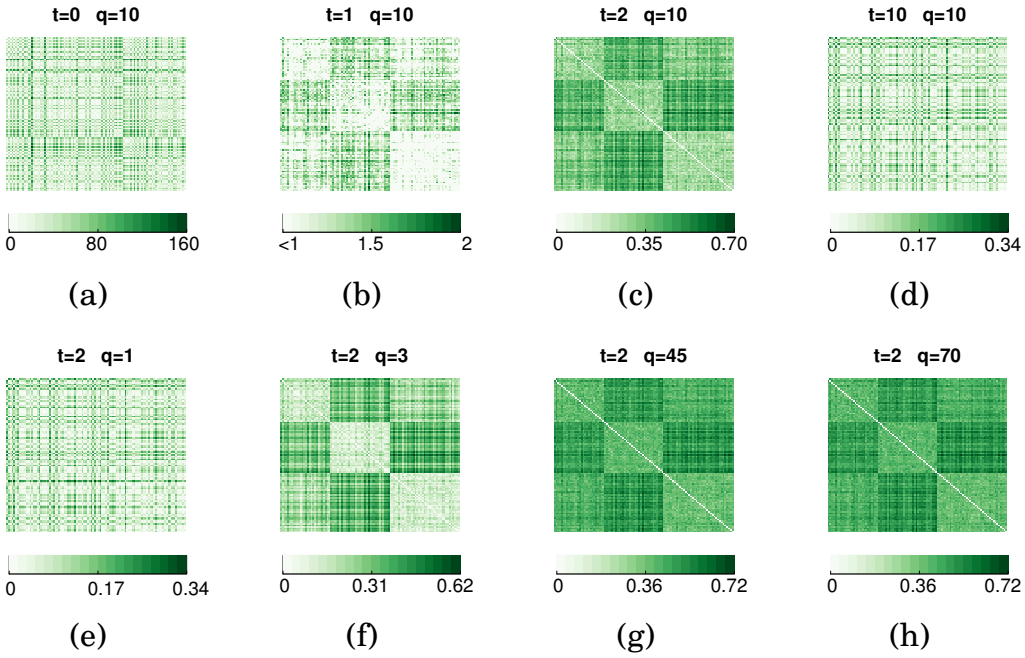


Figure 4.6: Generate a three-block adjacency matrix A by Equation 4.13 at $n = 100$, and compute the diffusion distances at each combination of (t, q). A visualization of adjacency matrix is provided in Figure 4.7 (a); upon fixing a good t, many choices of q preserve the block structure. Note that the first three elbows of eigenvalues are $(1, 45, 70)$ and $t^* = 2$, so panel (g) is the optimal diffusion map by DMGC.

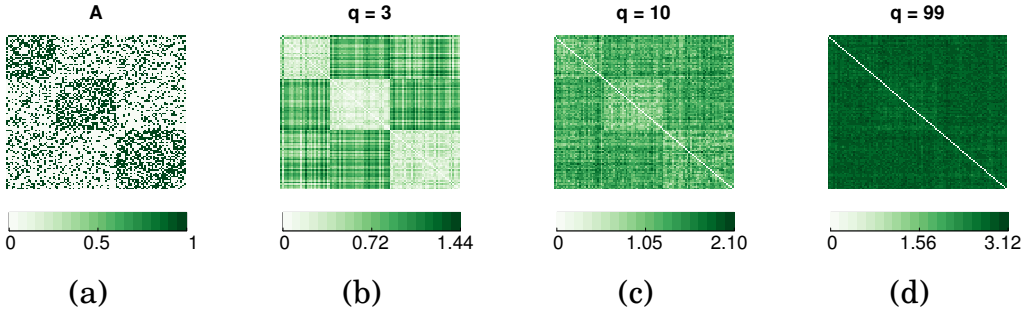
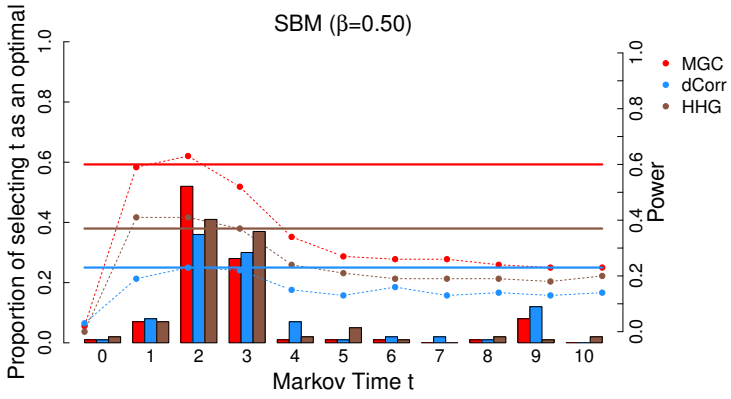


Figure 4.7: Panel (a) shows the adjacency matrix of three-block adjacency matrix A generated by Equation 4.13. Panel (b)–(d) show the Euclidean distance matrix of ASE at increasing q, using the same adjacency matrix of Panel (a). Only ASE at $q = 3$, namely at the correct dimension, is able to display a clear block structure. Note that the first three elbows are $(1, 45, 70)$, so ASE has a more obscure block structure when the dimension is chosen via the scree plot, comparing to the DMGC embedding in Figure 4.6 (g).

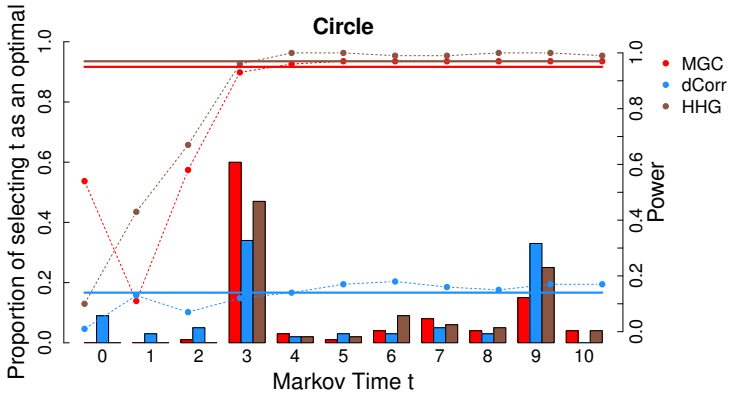
robust against q, e.g., Figure 4.6 (e)–(h) show that for a wide range of q the block structure is preserved in the resulting diffusion maps including the second elbow, so the DMGC embedding preserves the dependency structure well.

On the other hand Figure 4.7 shows that a choice of t without maximizing the dependency can be very sensitive to the choice of q, and may fail to preserve the dependency structure. Figure 4.7 shows the Euclidean distance of the adjacency spectral embedding (ASE) (Sussman et al., 2012) applied to the same adjacency matrix. For ASE, the correct dimensional choice equals the number of blocks, i.e., the distance matrix at $q = 3$ shows a clear block structure (Figure 4.7 (b)). However, a slight misspecification of q can cause the embedding to have a more obscure block structure, and the elbow method often fails to find the correct q for ASE.

Next we compare testing performance of the DMGC embedding $\mathbf{U}^{t^*}$ versus all other diffusion maps $\mathbf{U}^t$, e.g., both ASE and graph Laplacian embedding are equivalent to $\mathbf{U}^{t=1}$ up-to a linear transformation. Figure 4.8 shows the proportion of choosing t as the optimal among $\{0, 1, 2, \ldots, 10\}$ and the testing power for each t and also t^*. Figure 4.8 (a) illustrates that under the SBM dependency structure in Equation 4.14 with $\beta = 0.50$, diffusion MGC is mostly likely to choose $t^* = 2$ as the optimal time-step, and the testing power is almost equivalent to the best power among all $t \in \{0, 1, 2, \ldots, 10\}$. The same phenomena hold for diffusion DCORR and diffusion HHG, and Figure 4.8 (b) illustrates another



(a)



(b)

Figure 4.8: Testing power comparison between DMGC and MGC on each diffusion map. Using $m = 100$ replicates, the solid red line plots the power of $\text{MGC}_n^*(\{\mathbf{U}^t\}, \mathbf{X})$; the dash line plots the power of $\text{MGC}_n(\mathbf{U}^t, \mathbf{X})$ for $t \in \{0, 1, 2, \ldots, 10\}$; the bar plot shows the proportion that diffusion MGC selects each $t \in \{0, 1, 2, \ldots, 10\}$ as the optimal t^*. Diffusion HHG and diffusion DCORR are also added by different colors. For each method, the diffusion statistic is able to achieve an excellent power that is almost equivalent to the best possible power among all t, implying that the methodology is able to identify the graph embedding that best preserves the dependency structure.

RDPG simulation example by Equation 4.16.

Indeed, by utilizing a collection of $\{\mathbf{U}^t\}$ and identifying the strongest dependence signal, the smoothed maximum statistic has always achieved satis-

factory performance throughout the experiments in both the testing power and the resulting embedding, which does not rely on cross validation nor on multiple testing. On the other hand, most of the existing network methodology relies on a single embedding choice, therefore either falls short in practice due to a poor embedding choice or requires computationally intensive cross validation and further corrections to avoid potential over-fitting.

4.6 Real Data Application

As an illustrative example, we apply our distance-based tests on the neuronal network of hermaphrodite *Caenorhabditis elegans* (C.elegans) composed of 279 nonpharyngeal neurons connected each other through chemical and electrical synapses (Varshney et al., 2011). Each node represents an individual neuron and edge weights indicate the number of synapses between them. Among a few known attributes including types of neurotransmitter and role of neurons, we use one dimensional, continuous position of each neuron as a nodal attribute X. Figure 4.9 shows that neurons at low location and high location are connected to other neurons distributed throughout the region; while those at the relatively middle of location are connected to the neurons only within the narrower area. The independence test between synapse connectivity and each neuron's position can be connected to growing studying on relationship

between physical arrangement and functional connectivity in C.elegans (Chen et al., 2006; Kaiser and Hilgetag, 2006) or in others' (Cherniak et al., 2004; Alexander-Bloch et al., 2012). For the purpose of analysis we binarize and symmetrize both chemical and electrical synapses and add them together to simplify the adjacency matrix that represents overall synapse connectivity of C.elegans. We apply MGC, DCORR, HHG, and FH to testing independence between connectivity through synapses and neuron's position. All of these tests result very low p-values less than 0.002. Dimension of local distance corre-

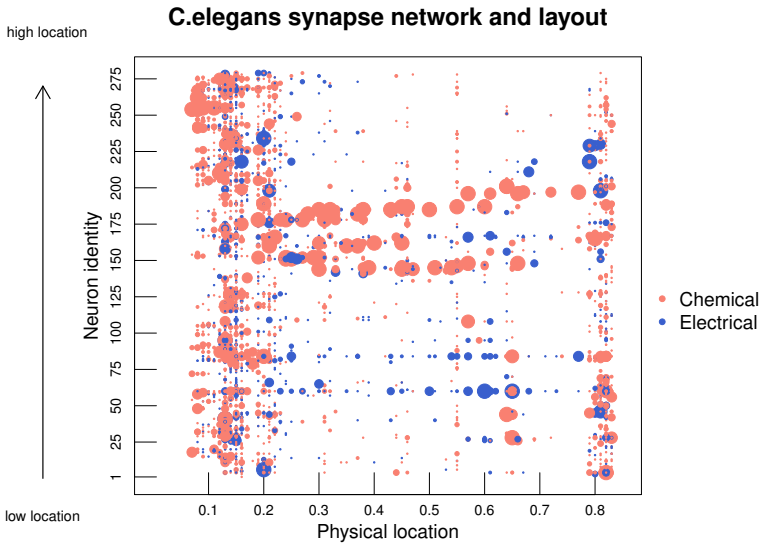


Figure 4.9: Each dot represents the existence of synapses from each neuron at y-axis indexed from low location to high location to other neuron on x-axis at certain position among 68 different locations. Color of dots represents synapse type, either chemical or electrical, and size of dots is proportional to the number of synapse but truncated at 10.

lation map ($\text{DCORR}^{kl}(\mathbf{U}, \mathbf{X})$) depends on the number of unique neighborhood scale with respect to distance in diffusion maps from graph (synapse connec-

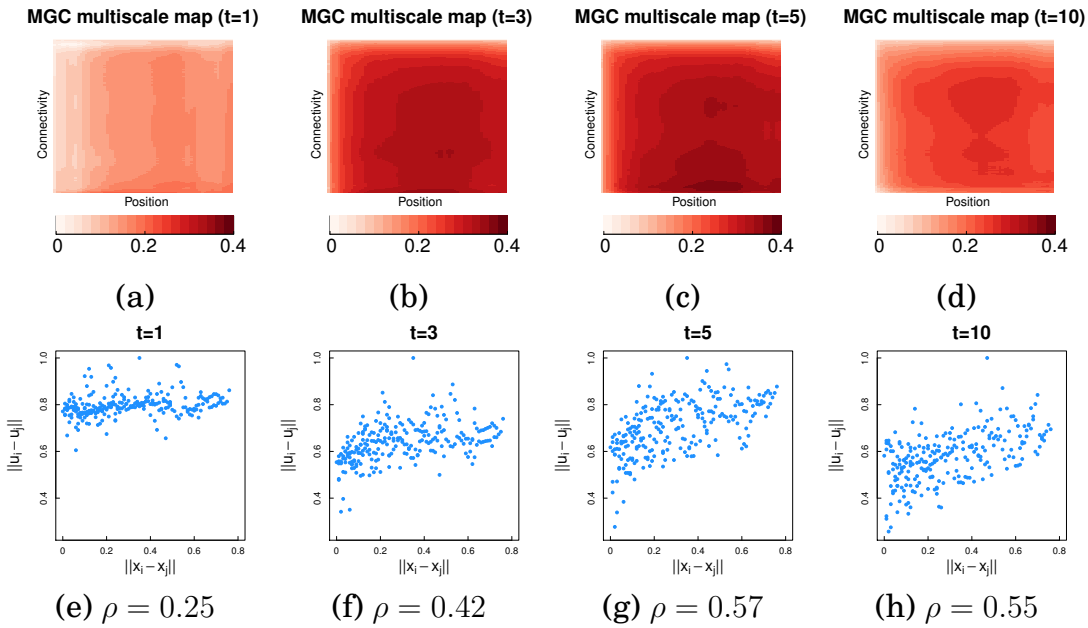


Figure 4.10: Local correlation maps of 279×68 matrices of which row and column specify the neighborhood scale in synapse' connectivity and position respectively. Figure (c) presents the correlation map at optimal time $t^* = 5$ where local optimality is achieved in local scale in position, which also implies that $\text{MGC}_n(\mathbf{U}^{t=4}, \mathbf{X})$, $\text{MGC}_n(\mathbf{U}^{t=5}, \mathbf{X})$, and $\text{MGC}_n(\mathbf{U}^{t=6}, \mathbf{X})$ result three maximum statistics among those of $t = 0, 1, 2, \ldots, 10$. Panel (e)-(h) show standardized Euclidean pairwise distance within $\{U_i^t\}$ scaled by its maximum according to Euclidean distance of $\{X_i\}$ for $t = 1, 3, 5, 10$, among of which correlation between two distances is most evident at $t^* = 5$ with highest correlation coefficient ρ.

tivity) and nodal attribute (position); here we have $\kappa = 279$ rows and $\gamma = 68$ columns each. Figure 4.10 presents local distance correlation map with respect to local distance of synapse connectivity ($\tilde{\mathbf{C}}^k(i, j)$) and local distance of position ($\tilde{\mathbf{D}}^l(i, j)$) across diffusion times. These plots show that the optimal local correlation is detected at non-global neighborhood choice in position, i.e. $l^* \neq 68$ (the global maximum), which provides evidence of non-linear dependence between connectivity and position; in addition, the plots show that when $t = 5$, this non-

global optimality manifests most, resulting DMGC statistic at optimal $t^* = 5$ where the optimal neighborhood choice is $(k^*, l^*) = (269, 42)$. Figure 4.10 (e)-(d) illustrate the rough relationship between Euclidean distance in diffusion maps and nodal attributes at different diffusion time ($t = 1, 3, 5, 10$). These present that correlations between two distances are most clear at $t^* = 5$. These findings support the results of DMGC along with the choice of optimal scales.

4.7 Discussion

Our contribution in this chapter is three-fold. First, we propose a new method for testing dependency on network data, which combines various state-of-the-art techniques from different domains into a valid, consistent, and interpretable test procedure that is also numerically superior. Second, the methodology in this chapter also defines a good correlation measure on nodes, thus enabling many popular statistical techniques on graph structure such as feature screening and outlier detection. Third, the utilization of diffusion maps not only warrants the integration with various types of distance-based correlations, but also makes the testing method robust against parameter misspecification. In these ways our procedure overcomes an important practical issue that often plagues existing approaches, and can provide an extremely useful tool for later inference tasks like classification and regression.

Nevertheless there are several follow-ups that would further advance the work. First of all, theoretical background in choosing smoothed optimal statistic is lacking; assuming t' is the true optimal diffusion time, it will be essential to find more systematic and reliable way to estimate t' and quantify variability in the estimated optimal t^*, based only on the testing statistics. This would possibly reduce computational burden instead of going over all possible diffusion times, e.g. $t = 0, 1, 2, \ldots, 10$. Moreover, even though we briefly discussed one example in Section 4.5, it is still obscure what is the impact of dimensional choice of q on diffusion map embedding and impact of combinational choice of (q, t) for diffusion map on testing. Therefore it is a natural next step to provide more efficient and rigorous grounds for choosing tuning parameters for multiscale test statistics. Finally, since one can apply diffusion to any graph, and one can think of any affinity (or kernel) matrix as a graph, this method can straightforwardly be applied to more general testing scenarios, which will be of interest for future work.

Acknowledgement

The authors thank Dr. Minh Tang and Dr. Daniel Sussman for their insightful suggestions to improve the paper. This work was partially supported by the National Science Foundation award DMS-1712947, and the Defense

CHAPTER 4. MULTIVARIATE NETWORK DEPENDENCE TESTING

Advanced Research Projects Agency's (DARPA) SIMPLEX program through SPAWAR contract N66001-15-C-4041.

Chapter 5

Identifying Causally Influential Subjects on a Social Network

Researchers across a wide array of disciplines are interested in identifying the most influential node(s) in a network. We argue that, although influence is often defined only implicitly in these literatures, the operative notion is inherently causal: influential nodes are those on which we would intervene in order to achieve the greatest effect across the entire network. We review existing measures of influence, which usually rely on features of the network structure or on simple diffusion models for the flow of information/outcomes over network nodes. We illustrate that popular measures of influence fail to capture true causal influence in general, and propose a class of new measures of nodal influence based on the strength of a causal effect of an intervention on cer-

tain characteristics of the node(s) on the outcomes observed across the entire network. We illustrate estimation of influence using data on Supreme Court justices' decisions.

This is a joint work in collaboration with Elizabeth Ogburn and Ilya Shpitser.

5.1 Introduction

Networks are collections of nodes, which represent entities such as people, institutions, genes, or brain regions; ties between pairs of nodes represent various forms of connections between them (Newman, 2018). For example, in a social network the ties may correspond to friendship, family, coworker, or neighbor relationships. The study of networks is booming in biology (Simko and Csermely, 2013; Wang et al., 2014), economics (Banerjee et al., 2013), statistics (Shalizi and Thomas, 2011; Smith et al., 2018), psychology (Robinaugh et al., 2016), physics (Albert and Barabási, 2002; Castellano et al., 2009), computer science (Kempe et al., 2003; Chen et al., 2009, 2010), and beyond. We are concerned with a commonly studied problem across all of these disciplines: identifying the most important or influential node(s) in a given network. This problem has implications for predicting outcomes or processes in a network, for designing interventions on a network, and for understanding

the dynamics underlying a network. Despite the vast literature on identifying important or influential nodes (Kempe et al., 2003; Borgatti, 2005; Fowler et al., 2007; Kitsak et al., 2010; Aral and Walker, 2012), few researchers have clearly defined "importance" or "influence," and those that have generally resort to model-dependent definitions that may not generalize beyond a particular mathematical model of network dynamics. We claim that in most, but not all, contexts, the desiderata for important or influential nodes correspond to a causal definition of influence: on which nodes should we intervene in order to have the greatest impact across the entire network?

Although the two terms are often used interchangeably in the existing literature (e.g. Lü et al. (2016)), we will distinguish between the overarching concept of *importance*, which may refer to predictive/descriptive or causal notions, and *influence*, which is an inherently causal notion representing one kind of importance. We will refer to notions of importance that do not correspond to influence as "descriptive importance" As examples of the latter, consider *PageRank*, which was originally designed to assess the relative importance of websites by counting the number of links to the websites across the web (Page et al., 1999), and *h-index*, which quantifies the importance of researchers by the number of citations their papers have received (Moed, 2006). PageRank is meant to capture the usefulness or desirability of a website and h-index the productivity and impact of a researcher's body of work. These are indeed purely descrip-

tive versions of importance (though of course a popular website or a productive researcher could also wield influence).

Measures of descriptive importance can be used to predict dynamics in a network, but they cannot generally be used to understand the mechanisms by which those dynamics operate or to predict the impact of interventions on the network; those endeavors require causal concepts, which are often of more interest to researchers. For example, researchers have attempted to identify influential nodes in social networks in order to learn how targeted advertising affects overall sales (Trusov et al., 2009; Katona et al., 2011), to stop the spread of disease through targeted vaccination efforts (Perisic and Bauch, 2009; Bauch and Galvani, 2013), and to maximize the diffusion of information across an entire network (Banerjee et al., 2013). Yet even when the goal is to find causally influential nodes, causal methods have been used only exceedingly rarely (Smith et al., 2018). Instead, researchers most often identify the most *central* nodes in a network, or posit a model for diffusion of information, behavior, or other outcomes over the network and define influence in terms of the parameters of the diffusion model. This discrepancy between measures of influence and the causal nature of the underlying research question may help explain the failure of some strategies for disseminating information or changing behavior via influential nodes to perform as expected (Paluck et al., 2016; Chin et al., 2018).

In what follows we focus on the identification of influential nodes in social networks, but the concepts we discuss can extend naturally to other kinds of networks (biological, institutional, etc.). In Section 5.2, we review the existing literature. In Section 5.3 we use concepts from causal inference to propose a new class of definitions of influence. In Section 5.4, we compare popular measures of influence to ours. Section 5.5 concludes.

5.2 Existing Measures of Influence

5.2.1 Preliminaries

We make the routine assumption that influence operates (only) through network ties. A network tie, represented by an *edge*, connects pairs of subjects and implies some kind of relationship between them. Each subject, or *node*, can exert influence on its peers and can also be susceptible to its adjacent peers influence via edges. Depending on the context, edges can transmit information, political power, infectious disease, or gossip, or can induce collaboration or shared behavior. Ties can be directed, representing one-way relationships, or undirected, representing symmetric relationships; they can be binary, representing the presence or absence of a tie, or weighted, representing ties of different strengths. In what follows we focus on the simplest and most com-

mon setting of binary edges.

The structure of an n-node network is encoded in the $n \times n$ *adjacency matrix* $\mathbf{A}$, where $A_{ij} = 1$ if there is an edge from node i to node j and $A_{ij} = 0$ otherwise. If the network is undirected, $A_{ij} = A_{ji}$; while if an edge from node i to node j does not necessarily imply an edge from node j to node i, $\mathbf{G}$ is directed. Because some relationships may be asymmetric due to power differences or social dynamics, e.g. relationships between boss/employee, leader/follower, teacher/student, etc., we consider directed networks in this chapter. We denote a network by $\mathbf{G}$ and denote its set of nodes as $\mathbb{V}(\mathbf{G})$.

Much of the existing literature on influential nodes relies entirely on network structure as given by the adjacency matrix in order to define influence. Another set of popular approaches rely on models for the diffusion process of information or behavior over the network, which clearly plays an important role in addition to network structure. For example, an infectious disease may spread differently from gossip over the same network. Although node influence depends both on network structure and the diffusion process, and although most of the existing literature defines influence in terms of one or both of these features of the network, neither of these concepts explicitly defines influence (Smith et al., 2018).

5.2.2 Centrality measures of influence

Node importance has been widely measured using centrality with the implicit or explicit goal of measuring the influence of one node on the whole network (Kiss and Bichler, 2008; Bakshy et al., 2011; Chami et al., 2014).

Degree centrality is the most popular measure of node importance and is based only on the number of ties per node (Freeman, 1978). In a directed network, influence is usually measured by out-degree, or the number of edges emanating from each node to other nodes. Two other popular centrality measures are betweenness and closeness centrality. The betweenness of node v is defined as the sum of the proportions of the shortest paths between all pairs of nodes pass through the node v, and closeness of node v under a directed network is proportional to the inverse of the average length of the geodesic distance between the node v to all other nodes in G (Freeman, 1978). Finally, eigenvector centrality is defined through the eigenvector associated with the greatest eigenvalue of the adjacency matrix (Bonacich, 1987), and the eigenvector centrality of one node is proportional to those of its adjacent nodes. Many variants of eigenvector centrality have been proposed such as Katz centrality (Katz, 1953), PageRank (Page et al., 1999), and principal component centrality (Ilyas and Radha, 2011).

Using centrality measures as proxies for influence implicitly relies on specific diffusion models for the flow of information through edges. Borgatti (2005)

showed that different centrality measures will capture influence under different diffusion models. For example, both betweenness and closeness centrality presume that information travels between two nodes only through the shortest available path. If node v_1 exerts influence on v_2 via a higher frequency of information leaving v_1 and arriving at v_2, then betweenness centrality measures influence. But if, on the other hand, what matters is the time until the first arrival of information from v_1 to v_2 then closeness centrality is the operative measure of influence.

However, the implicit assumptions about the diffusion process (e.g. traveling only through the shortest paths) and the targeted outcomes (e.g. frequency of stopping while traveling or first arrival times) have rarely been specified or acknowledged in research that uses centrality measures to capture influence. Furthermore, when the relationships between nodes do not correspond to a small class of diffusion models, what researchers describe as influence is often far removed from the explicit notion of the centrality except in a very few cases (we will describe one such case in Section 5.4).

5.2.3 Influence defined through diffusion processes

Another popular approach to measuring influence in networks is to specify a particular diffusion process, the *threshold model* (Granovetter, 1978) or the *cascade model* (Goldenberg et al., 2001), and to identify influential nodes by

analyzing the process for a specific network of interest, usually via simulation (Kempe et al., 2003; Chen et al., 2010; Narayanam and Narahari, 2011). The consequent results are heavily dependent on the presumed diffusion models as well as on a correctly specified network structure. This method has been used to analyze infectious disease epidemics using standard epidemic models like the susceptible-infected-susceptible model (SIS model) or susceptible-infected-recovered model (SIR model) (Bailey et al., 1975) (e.g. Saito et al. (2012); Sikic et al. (2013)). The literature is full of other examples in which methods for determining influence associated with nodes are valid only under a particular diffusion process (e.g. Aral and Walker (2012); Beaman et al. (2015)).

Relatively recently, several researchers have defined new centrality measures based on different diffusion models in order to capture influence (Sikic et al., 2013; Banerjee et al., 2014; Saito et al., 2016). As an example, Banerjee et al. (2013) defined a *diffusion centrality* which requires stringent assumptions about the diffusion process–information flows from one node to its adjacent nodes with a fixed probability independently at each period of time, and this continues for a specified period of time. Diffusion centrality of node i can be interpreted as the expected number of times that information initiated from the node i reaches any other nodes during the specified period. Here, a targeted outcome is explicitly specified, but this interpretation is only valid when the assumed diffusion model is true.

When diffusion models are not correctly specified, influence measures based on those models fail to accurately predict the effect of interventions on the network. Therefore, influence measures depending on diffusion processes are not reliable estimands for influence unless researchers have explicit knowledge of how outcomes travel across network ties.

5.2.4 Influence in statistical mechanics

Statistical mechanics provides a probabilistic framework to understand macroscopic phenomena as a result of behaviors and interactions among microscopic constituents (Chandler, 1987; Bialek et al., 2012), and especially to understand thermodynamics (Gibbs, 2014). Many researchers have proposed statistical mechanics approaches to identifying influential nodes, usually based on attempts to describe social dynamics using models developed for thermodynamics (Bahr and Passerini, 1998; Albert and Barabási, 2002; Castellano et al., 2009; Bialek et al., 2012; Lucas, 2013), and especially for ferromagnetic interactions, as we describe below.

To illustrate, social dynamics have often been compared to ferromagnetic interactions in magnets (Castellano, 2012); where one atomic spin is dependent on others in microscopic perspective while the state among ferromagnets makes a transition from irregular to regular phase in macroscopic world. Researchers have used this phenomenon as an anology to an individual's deci-

sions that are influenced by those of others in social network. Thus just as the collective behavioral changes in spins are traditionally modeled by the Ising model (Binney et al., 1992), Ising model has also modeled social dynamics for collective behavior or opinion formation (Grabowski and Kosiński, 2006) in social network. As an example of Ising model for social influence study, Lee et al. (2015) proposed a maximum entropy model with a particular application to quantifying the influence of Supreme Court justices of the United States. In Lee et al. (2015) one of the proposed measures for influence associated with each Supreme Court justice is the impact of individual perturbations around his or her average vote into the resulting, collective outcomes, e.g., the majority vote. Liu et al. (2010) also used the Ising model to predict collective opinion formation in a network in order to search for the subset of nodes exhibiting the largest influence by pretending that votes of the subset of justices are set to be fixed. See Klemm et al. (2012); Lucas (2013); Lynn and Lee (2016) for more examples that apply the Ising model to understand collective behaviors and identify the influence of each subject in a network.

Statistical mechanics used in social influence study understands collective behavior in social network as behaviors of physical movements, and numerous studies described above provide a measure of influence based on this understanding. However, there is no guarantee that human's collective behaviors would well be fitted into the framework for molecular behaviors; for instance,

while thermodynamics are governed by *the laws of thermodynamics* (Callen, 1998), no particular laws regulate how human subjects behave individually and collectively, and no validation has been performed yet to justify the compatibility between social dynamics and thermodynamics. In Section 5.4 we use a variant of the Ising model to identify the influence of Supreme Court justices based on the justification of its use under certain conditions (Ogburn et al., 2018a).

5.3 Identifying Causally Influential Nodes

Most of the research described above is motivated by the problem of maximizing (or minimizing) the chance of observing certain collective outcomes, or population-level changes, via the least intensive intervention (Ballester et al., 2006; Klemm et al., 2012; Kim et al., 2015; Chin et al., 2018). To put this more concretely, an intervention on the most influential nodes has the largest *causal effect* on collective outcomes.

5.3.1 Causal inference

Before defining influence as a causal quantity, we first introduce *potential outcomes* (Rubin, 1974, 1977, 2005). Consider a variable for binary intervention Z, either 0 or 1, then a pair of potential outcomes for unit i under $Z = 0$ and

$Z = 1$ are $(Y_i(0), Y_i(1))$ respectively, representing the outcome that we would have observed for unit i if we could have intervened to set $Z_i = 0$ or $Z_i = 1$; generally $Y_i(z)$ denotes a response of unit i when its value of intervention variable Z is z. Traditionally, researchers make the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1990b) that the potential outcome of i is not be affected by other units' treatment (*no interference*), but when subjects in a social network interact with one another and affect one another's outcomes, this assumption must be relaxed, as we discuss below.

Under our network setting, we should first consider three components for causal inference: unit, treatment, and outcome. All the nodes in a network are units of study; they are all subject to influence from one another. A treatment or intervention acts as a source or trigger of influence, and changes in outcome represent the consequence of such an intervention (Valente, 2012). We discuss which intervention is considered in studying influence in Section 5.3.4.

5.3.2 Causal inference and social networks

In a social network setting, the SUTVA assumption is often violated due to the possibility of a causal effect of one's treatment assignment on others' outcomes. The effect of one unit's treatment on another's outcome is known as *interference* (Rubin, 1990a; Sobel, 2006; Hudgens and Halloran, 2008; Tchetgen and VanderWeele, 2012; Aronow and Samii, 2013; Athey et al., 2018). For

example, vaccinating one unit not only decreases the risk of infection for that unit but also has a causal effect in decreasing the risk of infection for others, too (VanderWeele et al., 2012; Perez-Heydrich et al., 2014). Under interference, one unit's potential outcome could vary depending on others' treatment assignment; therefore, we must define potential outcomes with respect to the treatment assignments of all subjects, or at least of subjects within the sphere of influence (Rubin, 1990a; Hudgens and Halloran, 2008; Tchetgen and VanderWeele, 2012; Manski, 2013).

Suppose that there are N nodes in the underlying network G. Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)$ denote random variables for a node-level intervention (as defined in Section 5.3.4), and assume that these interventions are binary, either 0 or 1, for simplicity. Following the potential outcome framework of causal inference, denote the potential outcome of node i given a vector of treatment assignment of whole N nodes, $\mathbf{z} = (z_1, z_2, \ldots, z_N)$ as $Y_i(\mathbf{z})$, which would be node i's response if N nodes, including node i, on the network were assigned to z. This notation implies that the potential outcome of node i may vary depending on other nodes' intervention assignment. That is, even if $z_i = z_i'$, $Y_i(\mathbf{z})$ is not necessarily the same as $Y_i(\mathbf{z}')$ for any $i = 1, 2, \ldots, N$.

Depending on context, an intervention on influential nodes will maximize (e.g., increase the number of people who buy a new product) or minimize (e.g., reduce the number of people infected with a disease) the intervention's ef-

fect. Without loss of generality, let us assume that we want to maximize an average of potential outcomes in this chapter. Then the problem of identifying the most influential nodes can be translated into the problem of identifying the intervention assignment $\mathbf{z}$ such that expectation of average of potential outcomes under $\mathbf{z}$, $\sum_{i=1}^{N} E\left[Y_i(\mathbf{z})\right]/N$, has the largest value among a set of potential outcomes, $\{E[Y_i(\mathbf{z})] : i = 1, 2, \ldots, N, \text{ for all } \mathbf{z} \in \{0,1\}^N\}$, i.e., $\mathbf{z} = \text{argmax}_{\mathbf{Z}\in\{0,1\}^N} \sum_{i=1}^{N} E\left[Y_i(\mathbf{Z})\right]/N$. The number of intervened nodes is often fixed, most often at a single node to identify the single most influential node in the network.

Identification of the causal effect of intervention assignment requires a *no unmeasured confounding assumption* (Rubin, 1974), and for simplicity we consider a network experiment (Centola, 2010; Bond et al., 2012; Aronow and Samii, 2013; Aral and Walker, 2014; Paluck et al., 2016) where random assignment of the intervention is marginally independent of the potential outcomes (*network ignorability* (VanderWeele, 2008; Tchetgen et al., 2017)),

$$\mathbf{Z} \perp Y(\mathbf{z}) \text{ for all } \mathbf{z} \in \{0,1\}^N. \tag{5.1}$$

Identification and estimation of causal effects under this assumption has been discussed in the network experiment setting (Fowler and Christakis, 2010; Bond et al., 2012; Aral and Walker, 2012, 2014; Kim et al., 2015) where network

ignorability is guaranteed.

In observational settings network ignorability is likely violated, e.g. by confounding due to *homophily* or *assortative mixing*, in which similarities in outcomes create an edge between the nodes, as well as confounding due to many other latent factors (Aral et al., 2009; Shalizi and Thomas, 2011). However, causal effects are still identified if observed confounders suffice to render treatments and potential outcomes conditionally independent.

$$\mathbf{Z} \perp Y(\mathbf{z}) \mid \mathbf{C} \text{ for all } \mathbf{z} \in \{0,1\}^N, \tag{5.2}$$

for any observed confounders $\mathbf{C} = (\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_N)$ (*conditional network ignorability*) (Tchetgen et al., 2017).

5.3.3 A causal measure of influence

In connection with the causal nature of the underlying research question in identifying influential subjects, we define a function of influence associated with any nodes V, $\tau : V \to \mathbb{R}$, using counterfactual outcomes. For any $V \in \mathbb{V}(\mathbf{G})$:

$$\begin{aligned} \tau(V) &= \sum_{i=1}^{N} E\left[Y_i(\mathbf{z}_V)\right]/N \\ &= \sum_{i=1}^{N} E\left[Y_i(z_{j,j\in V} = 1,\ z_{k,k\notin V} = 0)\right]/N. \end{aligned} \tag{5.3}$$

If Y is a binary outcome indicating being active ($Y = 1$) or not ($Y = 0$), $\tau(V)$ denotes the proportion of active nodes over the network when we only intervene on a set of nodes $V \in \mathbb{V}(\mathbf{G})$; if Y is a continuous outcome, $\tau(V)$ is an average of n potential outcomes under the same intervention assignment of $\mathbf{z}_V = \{z_{j,j\in V} = 1,\ z_{k,k\notin V} = 0\}$. This influence measure of τ leaves the diffusion process between the intervention assignment and the responses unspecified. For instance, intervention of $\mathbf{z}_V$ may directly increase the outcome $Y_j,\ j \notin V$ (direct interference); or increase in Y_i ($i \in V$) due to the direct effect from $\mathbf{z}_V$ may change the outcome of $j,\ j \notin V$ (social contagion). Therefore, no matter how $\mathbf{z}_V$ affects the collective outcomes, $\tau(V)$ estimates the causal effect of intervening node(s) of V on the collective outcomes, and differences in $\tau(\cdot)$'s can be explained only through the intervention assignment under network ignorability. In the longitudinal setting, the treatment assignment might change the underlying network structures (e.g. Rand et al. (2011)) as well as the potential outcomes, and treatments can be time-dependent. In these settings, it is natural and reasonable to consider $\mathbf{z}_V$ in Equation 5.3 as an initial treatment assignment on node V so that influence from previous treatment assignments cannot be involved; in a similar manner, it is most relevant to consider stabilized outcomes at final time point as collective outcomes of interest among the evolving outcomes so that we allow enough time for influence to pass through the nodes in a network. Our definition of influence can handle the fact that

network topology might change over time.

Our measure of influence $\tau(\cdot)$ (Equation 5.3) is closely related to the influence function of $\sigma(\cdot)$ in Kempe et al. (2003) except that Kempe et al. (2003) only considered a binary outcome and direct intervention on the outcome without a formal causal statement. Their function of $\sigma(A)$ represents "the expected size of the activated set if A is targeted for initial activation"; when Y is binary, $\sigma(V) = N \cdot \tau(V)$ under the assumption of network ignorability when the intervention directly performs on the dynamic outcome of interest. Banerjee et al. (2013) also defined a notion of diffusion centrality as "the fraction of other units (households) who would eventually participate if this unit (household) were the only one initially informed." In this case targeted outcomes (participation) and interventions (being informed) are well defined, but effect of the interventions on the outcomes is not necessarily causal. Even if it were, diffusion centrality matches the influence measure of τ only if their diffusion model is correctly specified and influence is measured on a single node.

On the other hand, we may want to consider a general form of node influence as a function of other nodes' treatment assignment; for instance, in case we cannot control the treatment assignment for some nodes, changes in collective behavior due solely to the target nodes, e.g. $V \in \mathbb{V}(\mathbf{G})$, under a fixed treatment assignment on $\mathbb{V}(\mathbf{G}) \setminus V$ can serve as a measure for influence of V. Equation 5.4 generalizes $\tau(V)$ in Equation 5.3 such that $\tau(V; \mathbf{z}')$ may vary depending on the

treatment vector $\mathbf{z}' \in \{0,1\}^N$. Fixing $\mathbf{z}'$ to $\mathbf{0}_N$ yields the same results as $\tau(V)$ in Equation 5.3 subject to shifting by $\sum_{i=1}^{N} E\left[Y_i(\mathbf{0}_N)\right]/N$, where $\mathbf{0}_m$ denotes a vector of m zeros.

$$\tau(V;\mathbf{z}') = \sum_{i=1}^{N} \left\{ E\left[Y_i(z_{j,j\in V}=1,\ z_{k,k\notin V}=z'_{k,k\notin V})\right] - E\left[Y_i(z_{j,j\in V}=0,\ z_{k,k\notin V}=z'_{k,k\notin V})\right] \right\}/N. \tag{5.4}$$

The general influence measure of $\tau(V;\mathbf{z}')$ as a causal effect has already been discussed in Smith et al. (2018) where only the influence of a single node ($|V|=1$) is considered. This general definition of influence is useful given a single observation of a network where we often do not have control over the observed intervention; yet in identifying the most influential nodes independently of the observed intervention, forcing non-target nodes as control (as Equation 5.3) provides a fair comparison between the causal effects from different sets of nodes.

Sometimes very specific quantities other than τ might be of interest in studying influential nodes, e.g., locally defined influence originating from a particular node (Bond et al., 2012) or influence transferring between a particular pair (Fowler and Christakis, 2010). Even though these do not necessarily identify the most influential nodes in general, we can approach these problems using counterfactual outcomes by defining δ_{ij} as the influence of node i on node

j:

$$\delta_{ij} = E\left[Y_j(z_i = 1, \mathbf{z}_{-i} = \mathbf{0}_{N-1})\right] - E\left[Y_j(\mathbf{z} = \mathbf{0}_N)\right], \tag{5.5}$$

where $\mathbf{z}_{-i} = \mathbf{z} \setminus \{z_i\}$. Note that δ_{ij} does not merely denote the causal effect of i's treatment on j's outcome, but a set of all N treatments $\mathbf{z}' = \{z_i = 1, \mathbf{z}_{-i} = \mathbf{0}_{N-1}\}$ on j's outcome. Taken together, influence as the strength of causal effect of the intervention is coherent with the underlying research question, even though details of the measure may vary depending on the specifics of the research question.

5.3.4 Intervention as a trigger of influence

When measuring influence over a network, an intervention or treatment on the nodes can either be a direct intervention performed on the evolving outcome (e.g. increasing one-pound weight gain of each student at initial time point of study compared to that at previous reference time point to see who has the largest influence over all classmates' weights at the end of study) or an intervention through an external factor (e.g. giving a vaccine to each subject to see whose vaccination would prevent infectious diseases most efficiently). Liu et al. (2010) called the direct intervention on outcomes the *placement* of a fixed number of positive *seed*s; according to their definition, the placement of fixed

positive outcome values on the most influential nodes would yield the largest expected number of positive nodes in the network. Similarly Lee et al. (2015) also defined the influence of each Supreme Court justice as the impact of the fluctuations on his or her own outcome (small increments in tendency to vote in a certain way) on the collective vote results. When we are interested in direct interventions on units' outcomes, randomized experiments may be impossible, because it is often infeasible or unethical to randomize changes to the outcome directly. Nevertheless we can define the influence of a node (or nodes) as the causal effect of intervening the time-evolving outcome at baseline ($t = 0$) on the whole outcome at the final time point ($t = T$).

On the other hand, external factors, called *network intervention*, can be understood as "purposeful efforts to generate social influence" (Valente, 2012). Examples where external interventions generate social influence over collective outcomes can be found in the literature as well: information on a microfinance loan program (intervention) was injected by the microfinance institution to increase adoption rate of the program (outcome) in Banerjee et al. (2013); in a Facebook experiment, a Facebook message (intervention) about new products is randomly sent to a users Facebook friends and these recipients' adoption of the products was measured to identify each user's influence (Aral and Walker, 2012); and in Nickerson (2008), voters were given different face-to-face messages (intervention) to investigate their messages effects on the subjects family

members' propensity to vote (outcome). More examples of external interventions cascading in a social network can be found in Kim et al. (2015); Cai et al. (2015) and Paluck et al. (2016).

5.4 Simulations

The first section of the simulation investigates whether popular centrality measures agree with the influence measure of τ when only a single node is intervened upon, and the second section introduces a hypothetical experiment for identifying the most influential Supreme Court justices. Throughout the simulations, we assume network ignorability without any confounders. Details of the simulations can be found in the supplementary material. An implementing software `R` package is provided in `https://github.com/youjin1207/netchain`.

5.4.1 Agreement between centrality and influence

We consider out-degree, betweenness, closeness, eigenvector, and diffusion centrality for comparison with our influence measure τ under five different diffusion processes. As a measure of agreement between τ and each centrality

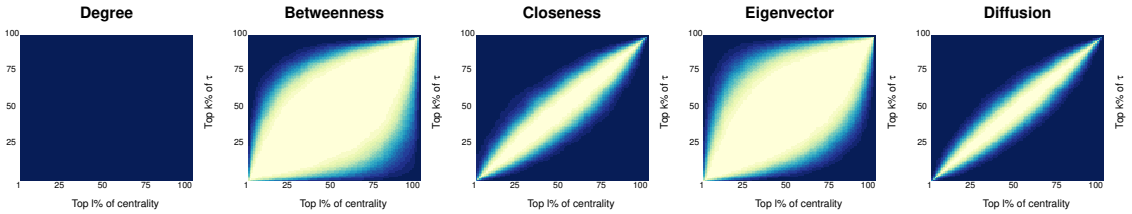
measure, we use Spearmans rank correlation ρ (Spearman, 1904) which quantifies correlations between ranks with respect to τ and with respect to centrality c. The Spearmans rank correlation ranges from -1 to +1 where +1 indicates a perfect positive monotonicity between τ and c. We assume five different dif-

	Out-degree	Betweenness	Closeness	Eigenvector	Diffusion
Homogeneous direct interference	1.00 (0.00)	0.63 (0.06)	0.94 (0.01)	0.66 (0.06)	0.96 (0.01)
Contagion process	0.92 (0.02)	0.67 (0.05)	0.89 (0.03)	0.59 (0.07)	0.90 (0.03)
Distance-dependent process	0.97 (0.01)	0.63 (0.06)	0.99 (0.01)	0.66 (0.06)	0.98 (0.00)
Traffic-dependent process	0.63 (0.06)	1.00 (0.00)	0.62 (0.06)	0.85 (0.03)	0.63 (0.06)
Homogeneous diffusion process	0.93 (0.01)	0.62 (0.06)	0.97 (0.01)	0.66 (0.06)	1.00 (0.00)

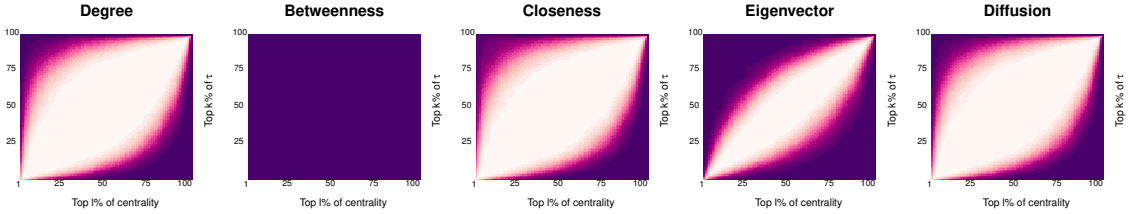
Table 5.1: Average of Spearman rank correlations and its standard errors between τ and c base on $r = 500$ independent replicates.

fusion models given an underlying social network comprised of $N = 100$ nodes, generate $r = 500$ Monte-Carlo replications for each model, and calculate the rank correlation of ρ between c and τ for each replication. Homogeneous direct interference implies a homogeneous causal effect of adjacent peers' intervention on the outcome; the contagion process implies a causal effect of an adjacent peer's outcomes on one's outcome over time; the distance-dependent process means a causal effect of others' intervention which depends on the geodesic distance between the nodes. The traffic-dependent process and homogeneous diffusion process are derived to match betweenness and diffusion centrality respectively. Even though at most one measure perfectly agrees with τ under each scenario, all five diffusion processes are based on stringent assumptions, e.g. homogeneous diffusion rate over nodes, interference only through adjacent peers, diffusion process through geodesic distance, etc. Details of each process

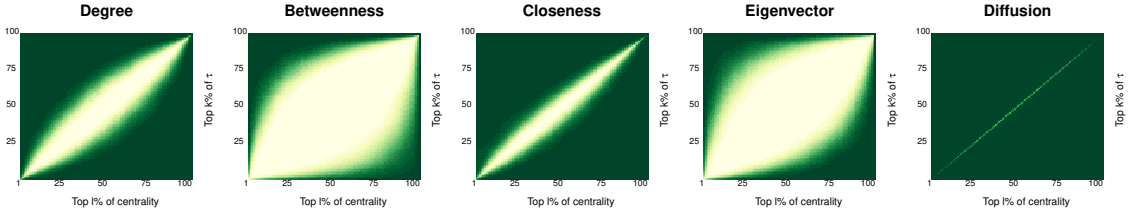
are illustrated in the supplementary material. The average of the rank correlations for each scenario is presented in Table 5.1.



(a) Homogeneous direct interference



(b) Traffic-dependent process



(c) Homogeneous diffusion process

Figure 5.1: Each matrix contains 100×100 cells; each cell illustrates how much the top $l\%$ of *influential* nodes from each centrality measure cover the top $k\%$ of *influential* nodes in terms of τ if $l \geq k$ (lower right corner); when $l < k$ (upper left corner) each cell represents the probability of how much the top $k\%$ in each centrality covers the top $l\%$ in τ. Under homogeneous direct interference (Figure 5.1a), degree centrality and τ are perfectly monotonic; the same is true for betweenness and τ in Figure 5.1b and for diffusion centrality and τ in Figure 5.1c.

Figure 5.1 illustrates the probabilities of interest like *"What are the chances of having the top* 10% *of nodes in term of* τ *in the top* 20% *of out-degree nodes?"*; under each of five diffusion processes we calculated agreement between cen-

trality and the influence measure τ empirically using $r = 500$ replicates. Other relevant figures for different diffusion processes can be found in the supplementary material. Even though almost all of the commonly used centralities fail to capture τ, under the specific data generating process of stringent assumptions, higher centrality implies higher influence of τ. For example, when the treatment assigned to each node has a homogeneous causal effect on its adjacent nodes and with other additional assumptions, higher out-degree centrality exactly implies higher influence, and vice versa; under the traffic-dependent process higher betweenness strictly implies higher influence (see proof in the supplementary material). However, given that these agreements require implausible conditions for most of the applications, all of the suggested centrality measures fail to identify causally influential nodes in general.

5.4.2 Influential nodes under latent confounding

Identifying the influence of nodes fails not only because of misspecified diffusion processes but also because of latent confounding between the intervention and outcome of interest. Through simple, empirical examples, we illustrate two cases where the measure of influence becomes useless due to confounding by latent variables. We present simplified numerical examples with

three nodes followed by specific illustrative descriptions. Confounding commonly occurs when a latent variable is highly predictable for an intervention variable and outcome variable at the same time. In Equation 5.6, consider a latent variable L denoting an indicator for joining a gym, and an intervention variable Z indicating recent weight loss. Assume that we are interested in whose weight loss is more influential in terms of making three people exercise. Let Y be an indicator for exercising or not, and pretend that there is no causal effect of Z on Y directly nor indirectly from peers. Instead, if adjacent peers join the gym ($L_j = 1$ for adjacent peer $j \neq i$), each person is more likely to exercise (positive effect of $A_{ij}L_j$).

$$
\begin{aligned}
L_i &\overset{i.i.d.}{\sim} Bernoulli(0.5),\ i = 1, 2, 3 \\
Z_i \mid L_i &\sim Bernoulli(0.5 + \beta(2L_i - 1)) \\
Y_i \mid \mathbf{L} &\sim Bernoulli(0.3 + 0.3L_i + 0.4 \sum_{j=1}^{N} A_{ij}L_j/N)
\end{aligned}
\tag{5.6}
$$

By varying a value of $\beta = (0.0, 0.1, 0.2, 0.3, 0.4, 0.5)$, we estimated false influence $\tau^*(1), \tau^*(2),$ and $\tau^*(3)$ ignoring the existence of a latent variable of L by averaging 10000 observations of (Y_1, Y_2, Y_3) under intervention of $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\} = \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$. Under the assumption of network ignorability and consistency of those average values, say $\overline{\hat{\tau}^*}$, would be close to the true value

of τ's. Because $\mathbf{Z} = (Z_1, Z_2, Z_3)$ has no causal effect at all on the outcome $\mathbf{Y} = (Y_1, Y_2, Y_3)$, $Y_i(\mathbf{z}) = Y_i(\mathbf{z}')$ for all possible values of $\mathbf{z} \neq \mathbf{z}'$ and for all $i = 1, 2, 3$, the influence measure $\tau(i)$ should be identical across $i = 1, 2, 3$.

β	$\overline{\hat{\tau}^*(1)}$	$\overline{\hat{\tau}^*(2)}$	$\overline{\hat{\tau}^*(3)}$
0.0	0.5336	0.5382	0.5369
0.1	0.5187	0.5277	0.5263
0.2	0.4952	0.5220	0.5050
0.3	0.4826	0.5095	0.4831
0.4	0.4642	0.5004	0.4639
0.5	0.4452	0.4858	0.4464

Table 5.2: Let τ^* denote the false influence measure ignoring a latent variable L. Each of $\hat{\tau}^*$ is derived by averaging three outcomes under $\mathbf{z}_1, \mathbf{z}_2$, and $\mathbf{z}_3$, and then we randomly generate potential outcomes 1000 times to calculate $\overline{\hat{\tau}^*}$. Under $\beta = 0.0$, $\tau^* = \tau$.

In Table 5.2, unit 2 falsely looks most influential because if unit 1 (or unit 3) is the only one who lost weight, unit 2 and unit 3 (or unit 1 and unit 2) are more likely not to sign up for the gym as β increases. Because they are friends with each other ($A_{23} = A_{12} = 1$) they are less likely to exercise together due to smaller $\sum_{i=1}^{N} A_{ij} L_j$; while if unit 2 is only one who lost weight, unit 1 and unit 3 are more likely not to sign up for the gym, but because they are not friends with each other ($A_{13} = 0$), the adverse effect on exercising is less so decrease in $\overline{\hat{\tau}^*}(2)$ is less significant than other two as β increases.

Now assume that a latent variable L denotes a previous midterm exam score for three units; an intervention variable Z is a binary indicator of taking an advanced online class or not. As β increases, L is more predictable for Z

in the same direction. Assume that we are interested in the influence of each unit taking an advanced online class on the final exam grades of the whole class (on the three units in our case); let $Y = 1$ mean higher grades and $Y = 0$ mean lower grades at final exam compared to the midterm exam. As shown in Equation 5.7 there is not causal effect of Z on Y but instead we assume a peer effect among students who have similar grades at the midterm to improve their grades; students are encouraged to do better if they are surrounded by students with similar performance. We can consider a matrix $W_{ij} = A_{ij}\mathbf{I}(L_i = L_j)$ as an another adjacency matrix forming as a result of homophily, and probability of $Y_i = 1$ (improved grades at the final exam) increases proportionally to the number of homogeneous friends.

$$
\begin{aligned}
&L_i \overset{i.i.d.}{\sim} Bernoulli(0.5),\ i = 1, 2, \ldots, N \\
&Z_i \mid L_i \sim Bernoulli(0.5 + \beta(2L_i - 1)) \\
&Y_i \mid L_i \sim Bernoulli(0.3 + 0.3L_i + 0.4 \sum_{j=1}^{N} A_{ij}\mathbf{I}(L_j = L_i)/N)
\end{aligned}
\tag{5.7}
$$

β	$\overline{\hat{\tau}^*(1)}$	$\overline{\hat{\tau}^*(2)}$	$\overline{\hat{\tau}^*(3)}$
0.0	0.5357	0.5442	0.5426
0.1	0.5286	0.5257	0.5319
0.2	0.5176	0.5039	0.5225
0.3	0.5128	0.4777	0.5098
0.4	0.5004	0.4407	0.4944
0.5	0.4853	0.3994	0.4860

Table 5.3: Estimates for τ^* were derived similarly to those in Table 5.2.

Contrary to Table 5.2, unit 2 looks less influential in Table 5.3 because if unit 2 is only one who takes an advanced lecture, unit 1 and unit 3 are likely to have poor grades on the previous exam. Because unit 1 and 3 do not know each other, their grades do not benefit from peer effects.

Therefore as presented in both cases of Equation 5.6 and Equation 5.7, ignoring any latent factors in the causal pathway between Z and Y easily results in a misleading influence measure.

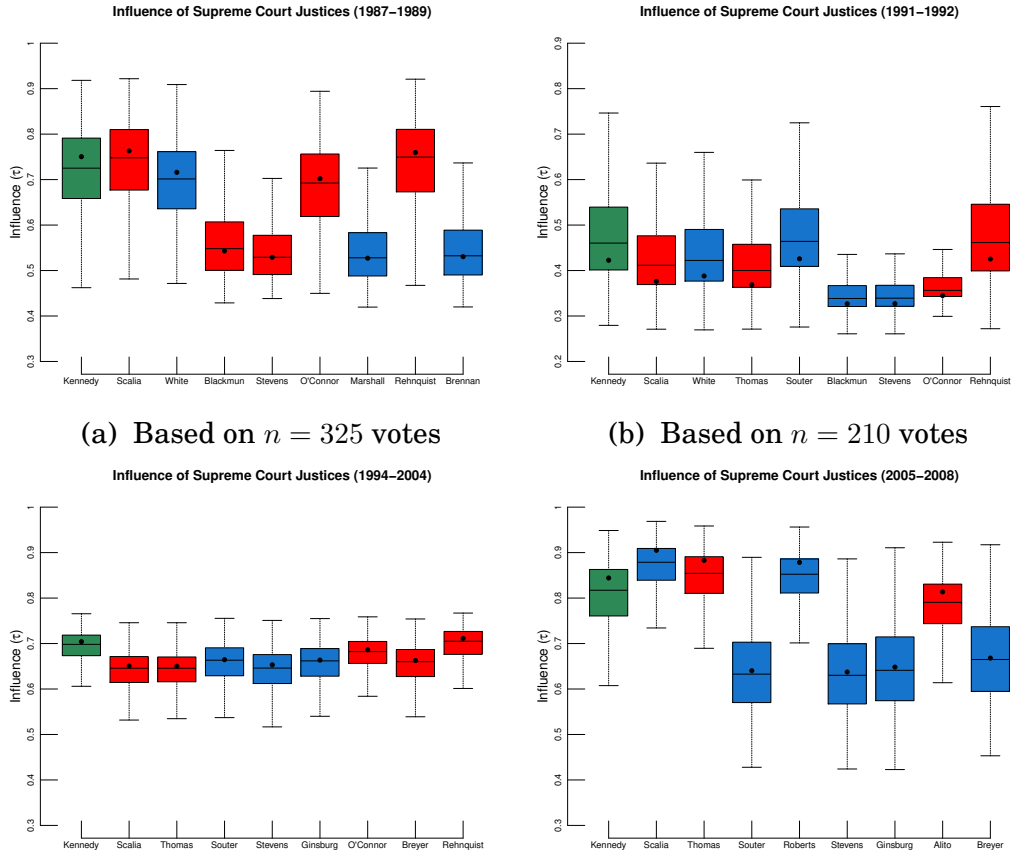
5.4.3 Identifying the most influential Supreme Court justice

In political science or general decision making, identifying whom to persuade or to whom additional information should be provided to elicit a certain social phenomenon are considered important issues (Huckfeldt and Sprague, 1995; Kenny, 1998). We introduce the example of the nine Supreme Court jus-

tices debating each other to reach a consensus. Identifying the most influential Supreme Court justices has been studied (Altfeld and Spaeth, 1984; Kosma, 1998; Pryor, 2017) using different definitions for *influence*. Recently, the influence of each justice has received a lot of attention after the retirement of Justice Kennedy. In our data analysis assuming certain hypothetical conditions, we investigate how Justice Kennedys influence varies depending on the other eight justices' votes and his relationship with them.

In this context, a binary random variable Y stands for the characteristics of each vote – liberal(1) or conservative(-1). We chose five periods having distinct sets of justices across time period for convenience, with more than 200 decisions made per period. We identify an edge between the justices, which implies some dependency between their votes. We then introduce a hypothetical justice-level treatment $\mathbf{Z}$ where $z_v = 1$ only increases justice v's chance of casting liberal votes with larger effects for justices with larger variability in their votes. Because each $\mathbf{Z}$ is randomly assigned to the justice, network ignorability is ensured in our simulation. To identify the causal effect of the treatment on the collective outcomes (e.g. the number of liberal votes or unanimous decisions), we assume a *chain graph* model (Ogburn et al., 2018a) on analytically identified edges by an ad hoc method of selecting pairs with significant two-way interaction effects from pairwise saturated log-linear model. In the chain graph model, to reflect the real voting tendencies of the justices,

we keep the estimates of the main effect and two-way interaction effects from the log-linear model using real data, and based on these coefficients we estimate the influence of each justice over five different periods. We assume no higher than pairwise interactions between the justices. Figure 5.2 illustrates a hypothetical influence with uncertainties from coefficients of the main effect and two-way interaction effect (box plot), showing that the influence of Justice Kennedy fluctuates over time (black dot). The influence of each justice in this particular case can be interpreted as the proportion of liberal votes among nine justices assuming that a hypothetical intervention that only increases the propensity to cast a liberal vote is assigned to each justice. It might also be of interest to study a pair of justices who have the largest influence. We can define τ as the probability of having unanimous decision, not an average liberal (or conservative) vote, under each treatment assignment. An important caveat here is that our results are valid only when the causal effect for the interventions on each justice we assumed are correct and when chain graph models are correctly specified. The chain graph we used for the inference also implies that the intervention of one justice does not have any direct effect on the other justices votes nor on any interaction between the justices and also assumes that contagion only occurs for pairs of justices specified in the model and, if any, no higher-order than two-way interactions exist. Given these assumptions and the hypothetical treatment effect, Figure 5.2 shows the absolute effects of each



(a) Based on $n = 325$ votes

(b) Based on $n = 210$ votes

(c) Based on $n = 893$ votes

(d) Based on $n = 252$ votes

Figure 5.2: Dots in each box plot denote the influence of each justice given main effects and two-way interaction effects from the log-linear model; while each box plot illustrates the empirical distribution of such influences based on coefficients from conditional log-linear models of the bootstrap sample. We have less variability in Figure 5.2c because there was smaller variability in each justices votes with large number of observations $(n = 893)$ so is causal effect of the treatment.

justice on the number of liberal votes vary across the court; for instance, providing each justice a hypothetical intervention to cast a liberal vote still leads to less than 50% of liberal votes on average from 1991 to 1992 while intervening any justice leads to greater than 60% of liberal votes on average from 1994 to 2004. Justice Kennedy's influence in terms of increasing liberal votes is rel-

atively higher than other liberal justices generally; this is probably because Justice Kennedy often sided with conservatives. His hypothetical influence is the highest among nine justices from 1994 to 2004, which implies that providing Justice Kennedy an incentive to make a liberal decision would maximize the number of liberal votes across the court compared to providing the same incentive to any of the other eight justices.

5.5 Discussion

In this chapter, we propose a class of causal measures of influence for nodes in a network. The research on measuring influence over social networks to date has tended to focus on specific features of networks and diffusion processes instead of defining target estimands without reference to a specific model. We found that most of the centrality measures that are dependent only on the network structure actually implicitly make highly stringent assumptions on diffusion processes and identify influential nodes only under these presumed diffusion process.

Above all, our main concern is the discrepancy between each of the estimators and what they originally were intended to measure. We suggest that no matter how plausible or practical centrality measures are to evaluate, the influence measure of τ as a causal effect is what we should instead target. Failure

to consider causal interpretation might lead to policy making based on spuriously 'influential' nodes, which would not achieve the anticipated effect on the network. Identification and estimation of causal effects under complex observational study is extremely challenging, and in this chapter we do not suggest a particular estimation method to identify influential nodes.

Targeting a parameter of interest, e.g. τ, rather than a model or an estimation method, allows flexibility but at the same time leaves unanswered the question of how to estimate this parameter. As we often observe a single network at a certain time point, rather than multiple, independent observations of the network, identification of influence for any set of nodes of interest is almost infeasible. Even if we had multiple observations, unless we are able to randomize every possible combination of treatment assignments, we should make some assumptions about the range of influence, which often requires some knowledge of network structure.

Model approximation using chain graph model (Ogburn et al., 2018a) might not be applicable for most of the network data. The lack of accurate knowledge about network data or model mis-specification is especially likely to engender bias. When the number of nodes is small enough (e.g., nine Supreme Court justices), and the number of observations is sufficiently large, a pair-wise saturated conditional log-linear model might be practical to use without any model assumptions except that of no higher than two-way interactions. However, this

still requires a substantial computational burden due to calculation of the normalizing constant (Besag, 1975).

Despite the technical difficulties of the estimation method, we suggest a new way to understand influence in a social network with coherent, causal interpretation. With this target estimand as a research objective, future research should focus on designing efficient and effective randomization schemes to infer influence of the nodes on a social network, and identification and estimation of τ in observational settings.

5.6 Appendix

5.6.1 Data generating models

Directed random graph $\mathbf{G} \sim \mathcal{G}(0.1, 0.05)$ was generated by `sample-sbm` function provided by `igraph R`, having two blocks with sample size of $N/2$ for each. Same-block probability is 0.1 and 0.05 otherwise. Denote the adjacency matrix by $\mathbf{A}$ of $\mathbf{G}$. For each diffusion process, graphs $\mathbf{G}$ are randomly generated $r = 500$ times.

The geodesic distance from node i to node j by `dist`(ij), a total number of the shortest paths from node i to node j by σ_{ij}, the total number of this shortest path passing through node v as $\sigma_{ij}(v)$. We assume network ignorability and

consistentcy conditions in this simulation so $E[Y_i|\mathbf{Z}] = E[Y_i(\mathbf{Z})]$ for all i.

1. Homogeneous direct interference

$$E[Y_i|\mathbf{Z}] = 0.1 + 0.3Z_i + 0.2\sum_{j=1}^{N} A_{ji}Z_j/N \tag{5.8}$$

2. Contagion process

$$E[Y_i^1|\mathbf{Z}] = 0.1 \tag{5.9}$$

$$E[Y_i^t|\mathbf{Z}] = 0.7E[Y_i^{t-1}|\mathbf{Z}] + 0.15Z_i + 0.1\sum_{j=1}^{N} A_{ji}Y_j^{t-1}/N,\ t = 2, 3, \ldots, 10. \tag{5.10}$$

3. Distance-dependent process

$$E[Y_i|\mathbf{Z}] = 0.1 + 0.3Z_i + 0.2\sum_{j=1}^{N} Z_j \texttt{dist}^{-1}(j, i) \tag{5.11}$$

4. Traffic-dependent process

$$E[Y_i|\mathbf{Z}] = 0.1 + 0.3Z_i + 0.5\sum_{j=1}^{N} Z_j \sum_{k=1, k\neq j\neq i}^{N} \sigma_{ki}(j)/\sigma_{ki} \tag{5.12}$$

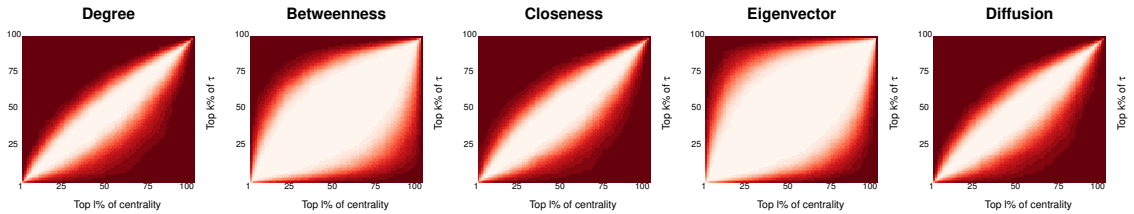
5. Homogeneous diffusion process

$$E[Y_i^0|\mathbf{Z}] = 0.3 \tag{5.13}$$

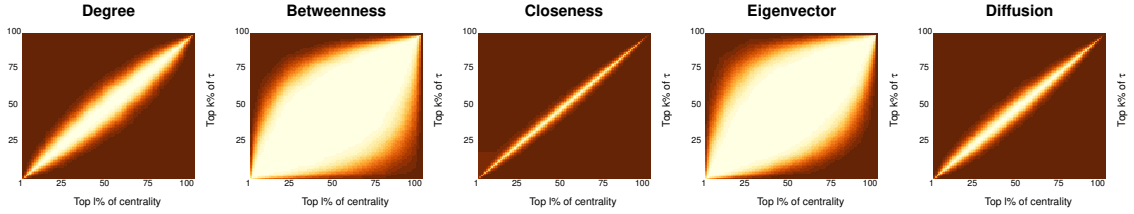
$$\text{For } t = 1, 2, 3, 4, 5: \tag{5.14}$$

$$E[Y_i^t|\mathbf{Z}] = 0.4E[Y_i^t|\mathbf{Z}] + \sum_{k=1}^{t}(0.3\mathbf{A})^k\mathbf{z} \tag{5.15}$$

We used diffusion centrality with diffusion rate of 0.1 under all processes except for homogeneous diffusion process where we specified true diffusion rate of 0.3 when deriving diffusion centrality. Figure 5.3 shows agreement matrices of five centrality measures under contagion process and distance-dependent process.



(a) Contagion process



(b) Distance-dependent process

Figure 5.3: Each matrix contains 100×100 cells of which each cell illustrates how much top $l\%$ of *influential* nodes in terms of each centrality measure covers top $k\%$ of *influential* nodes in terms of τ if $l \geq k$ (lower right corner); when $l < k$ (upper left corner) each cell represents the probability how much top $k\%$ in each centrality covers top $l\%$ in τ. Under contagion process (Figure 5.3a), degree, closeness, and diffusion centrality work reasonably well while both of betweenness and eigenvector centrality does not capture influence of τ well under two processes. Closeness centrality which is represented as a reciprocal of the sum of the geodesic distances between the node and all other nodes, agrees with τ under distance-dependent process but does not perfectly agree because τ is proportional to the sum of all reciprocal geodesic distances, not to the reciprocal of the sum.

5.6.2 Proofs

Denote a zero vector of length m by $\mathbf{0}_m$, i.e., $\mathbf{z} = \mathbf{0}_N$ is a null treatment assignment for all of N nodes. A vector of $\mathbf{z}_{-i}$ has a length of $N - 1$, removing an element of z_i from $\mathbf{z} = (z_1, z_2, \ldots, z_N)$.

Proposition 3. *Under network ignorability condition (Condition 5.1), when treating each node has a homogeneous effect only on its adjacent nodes and*

direct treatment effects as well as baseline distribution of potential outcomes under the null treatment ($\mathbf{z} = \mathbf{0}_N$) are homogeneous across all the nodes, higher out-degree centrality implies higher influence of τ, and vice versa.

Proof of Proposition 3. Assume that $\delta := E(Y_i(\mathbf{z} = \mathbf{0}_N)) > 0$, $\alpha := E(Y_i(z_i = 1, \mathbf{z}_{-i} = \mathbf{0}_{N-1})) - E(Y_i(\mathbf{z} = \mathbf{0}_N)) > 0$ for all $i = 1, 2, \ldots, N$, and $\beta := E(Y_i(z_j = 1, \mathbf{z}_{-j} = \mathbf{0}_{N-1})) - E(Y_i(\mathbf{z} = \mathbf{0}_N)) > 0$ for all edges from j to i. Then if out-degree of node u, denoted by $d_u (:= \sum_{i=1}^{N} A_{ui})$, is larger than out-degree of v, denoted by $d_v (:= \sum_{i=1}^{N} A_{vi})$, $\tau_u > \tau_v$.

$$
\begin{aligned}
\tau_u =& \sum_{i=1}^{N} E(Y_i(z_i = 1,\ \mathbf{z}_{-i} = \mathbf{0}_{N-1}))/N \\
=& \sum_{i=1}^{N} E(Y_i | z_i = 1,\ \mathbf{z}_{-i} = \mathbf{0}_{N-1})/N \\
=& \sum_{i=1}^{N} \left(\delta + \alpha z_i + \beta \sum_{k=1}^{N} A_{ki} z_k \right) /N \\
=& \delta + \alpha/N + \beta \sum_{i=1}^{N} A_{ui}/N \\
=& \delta + \alpha/N + \beta d_u/N \\
>& \delta + \alpha/N + \beta d_v/N \\
=& \tau_v
\end{aligned}
\tag{5.16}
$$

□

If $\tau_u > \tau_v$, we can easily show from the above equations that $d_u > d_v$.

Proposition 4. *Under network ignorability condition (Condition 5.1), if treating node k lying on shortest path from node $i(\neq k)$ to node $j(\neq k, i)$ has a homogeneous effect on node j with a size of $\sigma_{ij}(k)/\sigma_{ij}$ for all $i, j, k \in \{1, 2, \ldots, N\}$ and direct treatment effect as well as baseline distribution of potential outcomes under the null treatment ($\mathbf{z} = \mathbf{0}_N$) are homogeneous across all the nodes, higher betweenness centrality implies higher influence of τ, and vise versa.*

Proof of Proposition 4. Assume that $\delta := E(Y_i(\mathbf{z} = \mathbf{0}_N)) > 0$, $\alpha := E(Y_i(z_i = 1, \mathbf{z}_{-i} = \mathbf{0}_{N-1})) - E(Y_i(\mathbf{z} = \mathbf{0}_N)) > 0$ for all $i = 1, 2, \ldots, N$, and $\alpha := E(Y_i(z_j = 1, \mathbf{z}_{-j} = \mathbf{0}_{N-1})) - E(Y_i(\mathbf{z} = \mathbf{0}_N)) > 0$ for all edges from j to i. Then if betweenness of node u, denoted by $b_u (:= \sum_{i \neq u \neq j} \sigma_{ij}(u)/\sigma_{ij})$, is larger than betweenness of v, denoted by $b_v (:= \sum_{i \neq v \neq j} \sigma_{ij}(v)/\sigma_{ij})$, $\tau_u > \tau_v$.

$$
\begin{aligned}
\tau_u =& \sum_{i=1}^{N} E(Y_i(z_i = 1,\ \mathbf{z}_{-i} = \mathbf{0}_{N-1}))/N \\
=& \sum_{i=1}^{N} \left(\delta + \alpha z_i + \beta \sum_{k=1}^{N} z_k \sum_{j=1;\ j \neq k \neq i}^{N} \sigma_{ji}(k)/\sigma_{ji} \right) /N \\
=& \delta + \alpha/N + \beta \sum_{i=1}^{N} \sum_{j=1;\ j \neq u \neq i}^{N} \sigma_{ji}(u)/(\sigma_{ji} N) \\
=& \delta + \alpha/N + \beta b_u/N \\
>& \delta + \alpha/N + \beta b_v/N \\
=& \tau_v
\end{aligned}
\tag{5.17}
$$

□

5.6.3 Numerical experiment on Supreme Court justices

In hypothetical setting for numerical experiment, we introduce partial information on Supreme Court Justice data from Washington University Law Schools Supreme Court Database (`http://scdb.wustl.edu/ data.php`) (See Appendix B for details). For each of five different periods, we fitted an undirected edges among nine justice by ruling out insignificant interaction term from pair-wise saturated log-linear models. In order to reflect the magnitude of interactions between a pair of justices and justice-level propensity toward liberal (or conservative) opinion, we borrowed the estimated parameters in the log-linear model with only significant edges between node i and j, i.e. $e_{ij} = 1$, included in the model (Equation 5.18).

$$p\big(\mathbf{Y} = (y_1, y_2, \ldots, y_9)\big) = \frac{1}{B} \exp\left\{ \sum_{i=1}^{9} \alpha_i y_i + \sum_{i,j=1, e_{ij}=1}^{9} \beta_{ij} y_i y_j \right\}, \tag{5.18}$$

where B denotes a normalizing constant. We apply the maximum likelihood estimates of $\{(\alpha_i, \beta_{ij};\ i, j = 1, 2, \ldots, 9;\ e_{ij} = 1)\}$ from Equation 5.18 to the following simulation model of Equation 5.19.

$$p\big(\mathbf{Y} = (y_1, y_2, \ldots, y_9)|\mathbf{z}\big) = \frac{1}{B(\mathbf{z})} \exp\left\{ \sum_{i=1}^{9} \gamma_i z_i y_i + \sum_{i=1}^{9} \hat{\alpha}_i y_i + \sum_{i,j=1, e_{ij}=1}^{9} \hat{\beta}_{ij} y_i y_j \right\}, \tag{5.19}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_9)$ denotes hypothetical binary intervention of -1 or 1; γ_i represents the causal effect of intervention of z_i on y_i and we assume that γ_i is standard deviation of α_i from Equation 5.18 under the belief that justices who showed much variabilities in his or her votes they are likely to be influenced by intervention more. Since γ_i is always positive, intervention only increases the chance of liberal votes for all justices. Note, Equation 5.19 only includes up-to two-way interactions between the justices and does not include the interference term, e.g. $z_i y_j$ for $i \neq j$; hence estimated influence in Figure 5.2 is based on our own assumption of hypothetical, particular intervention.

Appendix A

Supplementary Material of Chapter 4

This is a joint work in collaboration with Chencheng Shen, Carey E. Priebe, and Joshua T. Vogelstein.

A.1 Proofs

Unless mentioned otherwise, throughout the proof section we always omit the superscript t for the diffusion map at a fixed t, i.e., we use $\mathbf{U} = \{U_i : i = 1, 2, \ldots, n\}$ instead of $\mathbf{U}^t = \{U_i^t : i = 1, 2, \ldots, n\}$ because most results hold for any t, similarly we use θ instead of θ^t whenever appropriate.

(Theorem 1). By the *de Finetti's Theorem* (Diaconis and Freedman, 1980; O'Neill,

2009; Orbanz and Roy, 2015), it suffices to prove that the diffusion map $\mathbf{U} = \{U_i : i = 1, \ldots, n\}$ is always exchangeable in distribution, i.e., for any n and all possible permutation σ, the permuted sequence $\mathbf{U}_\sigma = \{U_{\sigma(1)}, U_{\sigma(2)}, \ldots, U_{\sigma(n)}\}$ always distributes the same as the original sequence $\mathbf{U} = \{U_1, U_2, \ldots, U_n\}$.

Transforming Equation 4.3 in the main chapter into matrix notation yields

$$\mathbf{U} = \Lambda^t \Phi^T,$$

where $\mathbf{U}$ is the $q \times n$ matrix having U_i as its i^{th} column, $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ is the diagonal matrix having selected eigenvalues of $\mathbf{L}$, $\Phi = [\phi_1, \phi_2, \cdots, \phi_q]$ consists of the corresponding eigenvectors, $\cdot^t$ denotes t^{th} power, and $\cdot^T$ is the matrix transpose. It suffices to show that $\mathbf{U}$ and $\mathbf{U}\Pi$ are identically distributed for any permutation matrix Π of size n.

Given that the graph $\mathbf{G}$ is an induced subgraph of an infinitely exchangeable graph, it holds that $\mathbf{A}(\sigma(i), \sigma(j)) \overset{d}{=} \mathbf{A}(i, j)$, which further holds for the symmetric graph Laplacian $\mathbf{L}$:

$$\begin{aligned}
\mathbf{L}(\sigma(i), \sigma(j)) &= \mathbf{A}(\sigma(i), \sigma(j)) / \{\sum_j \mathbf{A}(\sigma(i), \sigma(j)) \sum_i \mathbf{A}(\sigma(i), \sigma(j))\}^{1/2} \\
&\overset{d}{=} \mathbf{A}(i, j) / \{\sum_j \mathbf{A}(i, j) \sum_i \mathbf{A}(i, j)\}^{1/2} \\
&= \mathbf{L}(i, j).
\end{aligned}$$

In matrix notation, $\Pi^T \mathbf{L} \Pi \overset{d}{=} L$ for any permutation matrix Π.

By eigen-decomposition, the first q eigenvalues and the corresponding eigenvector of $\Pi^T \mathbf{L} \Pi$ are Λ and $\Pi^T \Phi$, so it follows that at any t

$$\begin{aligned} & \Phi \overset{d}{=} \Pi^T \Phi \\ \Leftrightarrow \quad & \mathbf{U} = \Lambda^t \Phi^T \overset{d}{=} \Lambda^t \Phi^T \Pi = \mathbf{U} \Pi. \end{aligned}$$

Thus columns in $\mathbf{U}$ are exchangeable, i.e., the diffusion maps, $\{U_i \in \mathbb{R}^q : i = 1, 2, \ldots, n\}$, are infinitely exchangeable. By the *de Finetti's Theorem*, there exists an underlying variable θ distributed as the limiting empirical distribution, such that $U_i|\theta$ are asymptotically i.i.d. □

(Theorem 2). We first state three lemmas:

Lemma 1. *Under the same assumptions of Theorem 2, for any finite time-step t, the underlying distribution of U_i^t of the diffusion map is of finite first moment.*

Lemma 2. *The distance covariance of $(\mathbf{U}, \mathbf{X}) = \{(U_i, X_i) : i = 1, \ldots, n\}$ defined in Equation 4.5 in the Chapter 4 satisfies*

$$\text{DCOV}_n(\mathbf{U}, \mathbf{X}) = \int_{\mathbb{R}^{q+p}} |\hat{g}_{\mathbf{U},\mathbf{X}}(t, s) - \hat{g}_{\mathbf{U}}(t) \hat{g}_{\mathbf{X}}(s)|^2 dw(t, s), \tag{A.1}$$

where $w(t, s) \in \mathbb{R}^q \times \mathbb{R}^p$ is a nonnegative weight function that equals $(c_q c_p |t|_q^{1+q} |s|_p^{1+p})^{-1}$, c_q is a nonnegative constant, $\hat{g}_\cdot$ is the empirical characteristic function of $\{(U_i, X_i) :$

$i = 1, 2, ..., n\}$ *or the marginals, e.g.,* $\hat{g}_{\mathbf{U},\mathbf{X}}(t,s) = \frac{1}{n}\sum_{i=1}^{n} \exp\{\boldsymbol{i}\langle t, U_i\rangle + \boldsymbol{i}\langle s, X_i\rangle\}$ *with* $\boldsymbol{i}$ *representing the imaginary unit.*

Lemma 3. *Assume* $\mathbf{U} = \{U_i \sim F_U : i = 1, 2, \ldots, n\}$ *are conditional i.i.d. as* $U|\theta$*, and* $\mathbf{X} = \{X_i \overset{i.i.d.}{\sim} F_X : \ i = 1, 2, \ldots, n\}$*, and all distributions are both of finite first moment. It follows that*

$$\text{DCOV}_n(\mathbf{U}, \mathbf{X}) \to \text{DCOV}(U, X) \text{ as } n \to \infty,$$

where $\text{DCOV}(U, X) := \int_{\mathbb{R}^{q+p}} |g_{U,X}(t,s) - g_U(t)g_X(s)|^2 dw(t,s)$ *is the population distance covariance, and* $g_\cdot$ *is the characteristic function, i.e.,* $g_{U,X}(t,s) = E(\exp\{\boldsymbol{i}\langle t, U\rangle + \boldsymbol{i}\langle s, X\rangle\})$.

By Theorem 1, the diffusion maps U_i are asymptotically i.i.d. conditioned on θ, whose finite moment is guaranteed by Lemma 1. The nodal attributes X_i are i.i.d. as F_X of finite first moment as assumed in (C2). Therefore a direct application of Lemma 2 and Lemma 3 yields that

$$\text{DCOV}_n(\mathbf{U}, \mathbf{X}) \to \int_{\mathbb{R}^{q+p}} |g_{U,X}(t,s) - g_U(t)g_X(s)|^2 dw(t,s),$$

which equals 0 if and only if U is independent of X. As distance correlation is

just a normalized version of distance covariance, it further leads to

$$\text{DCORR}_n(\mathbf{U}, \mathbf{X}) \to c \geq 0, \tag{A.2}$$

for which the equality holds if and only if $F_{UX} = F_U F_X$. By Shen et al. (2018a), Equation A.2 also holds for MGC when it holds for DCORR. □

(Lemma 1). To prove that U is of finite first moment, it suffices to show that $\|U_i\|_2$ is always bounded for all $i \in [1, n]$.

By Equation 3, we have

$$\begin{aligned}
\|U_i\|_2^2 &= \sum_{j=1}^{q} \lambda_j^{2t} \phi_j^2(i) \\
&\leq \sum_{j=1}^{q} \lambda_j^{2t} \\
&\leq q,
\end{aligned}$$

where the second line follows by noting $\phi_j(i) \in [-1, 1]$ (the eigenvector ϕ_j is always of unit norm), and the third line follows by observing that $|\lambda_j| \leq \|L\|_\infty = 1$.

Therefore, all of U_i are bounded in ℓ_2 norm as $n \to \infty$, so the underlying variable U must be of finite first moment for any finite t. □

(Lemma 2). This lemma is a direct application of Theorem 1 in Székely et al.

(2007), which holds without any assumption on $(\mathbf{U}, \mathbf{X}) = \{(U_i, X_i) : i = 1, 2, ..., n\}$, e.g., it holds without assuming exchangeability, nor identically distributed, nor finite first moment. □

(Lemma 3). This lemma is equivalent to Theorem 2 in Székely et al. (2007), except the i.i.d. assumption is replaced by exchangeable assumption, i.e., the original set-up needs $(\mathbf{U}, \mathbf{X}) = \{(U_i, X_i) : i = 1, 2, \ldots, n\}$ to be independently identically distributed as F_{UX} with finite first moment; whereas the diffusion map $\{U_i : i = 1, 2, \ldots, n\}$ is asymptotically conditional i.i.d. with finite first moment.

Note that $\hat{g}_{\mathbf{U},\mathbf{X}}(t, s) = E(\hat{g}_{\mathbf{U},\mathbf{X}}(t, s)|\theta)$, and each term in $\hat{g}_{\mathbf{U},\mathbf{X}}(t, s)|\theta$ is asymptotically i.i.d. of each other. Thus

$$\begin{aligned}\int |\hat{g}_{\mathbf{U},\mathbf{X}}(t,s) - \hat{g}_{\mathbf{U}}(t)\hat{g}_{\mathbf{X}}(s)|^2 dw &= E(\int |\hat{g}_{\mathbf{U},\mathbf{X}}(t,s) - \hat{g}_{\mathbf{U}}(t)\hat{g}_{\mathbf{X}}(s)|^2 dw|\theta) \\ &\to E(\int |g_{U,X}(t,s) - g_U(t)g_X(s)|^2 dw|\theta) \\ &= \int |g_{U,X}(t,s) - g_U(t)g_X(s)|^2 dw,\end{aligned}$$

where the convergence in the second step follows from Theorem 2 in Székely et al. (2007) on the i.i.d. case. □

(Theorem 3). From Theorem 2, it holds that

$$\text{MGC}_n(\mathbf{U}^t, \mathbf{X}) \to c \geq 0 \tag{A.3}$$

for each t, with equality if and only if independence. The DMGC algorithm enforces that

$$\max\{\text{MGC}_n(\mathbf{U}^t, \mathbf{X}), t = 0, 1, \ldots, 10\} \geq \text{MGC}_n^*(\{\mathbf{U}^t\}, \mathbf{X}) \geq \text{MGC}_n(\mathbf{U}^{t=3}, \mathbf{X}),$$

thus Equation A.3 also holds when $\text{MGC}(\mathbf{U}^t, \mathbf{X})$ is replaced by $\text{MGC}^*(\{\mathbf{U}^t\}, \mathbf{X})$.

To show that the test is valid and consistent, it suffices to show that with probability approaching 1, $\text{MGC}_n(\mathbf{U}, \mathbf{X}_\sigma) \to 0$. This holds when $(U_i, X_i) \overset{i.i.d.}{\sim} F_{UX}$: the proof in supplementary of Shen et al. (2018a) shows that the percentage of partial derangement of finite sample size converges to 1 among all random permutations, such that with probability converging to 1 a permutation test breaks dependency.

For exchangeable $\{U_i\}$ here, we instead have $(U_i, X_i)|\theta \overset{i.i.d.}{\sim} F_{UX|\theta}$ asymptotically. The distribution of θ is the limiting empirical distribution of $\{U_i\}$, which is either asymptotically independent of all X_i or dependent only on finite number of X_i. Thus U_i is asymptotically conditionally independent with $X_{\sigma(i)}$ with

probability converging to 1, and we have

$$\text{MGC}_n(\mathbf{U}, \mathbf{X}_\sigma) = E(\text{MGC}_n(\mathbf{U}, \mathbf{X}_\sigma)|\theta) \to 0$$

Moreover, when the transformation from $\mathbf{A}$ to $\mathbf{U}^t$ is injective, we have

$$\begin{aligned} & A \text{ is independent of } X \\ \Leftrightarrow\ & U^t \text{ is independent of } X \text{ for all } t \\ \Leftrightarrow\ & \text{MGC}_n(\mathbf{U}^t, \mathbf{X}) \text{ is asymptotically } 0, \end{aligned}$$

where the second line follows from injective transformation, and the third line follows from Theorem 1 and Theorem 2. Thus DMGC is consistent between A and X.

Note that without the injective condition, the reverse direction of the second line may not always hold, i.e., when the diffusion maps are independent from the nodal attributes, the adjacency matrix may be still dependent with the nodal attributes. In that case, DMGC is still valid but the dependency may not be detected by DMGC. □

(Corollary 1). (1) Changing the test statistic only affects Theorem 2. Both DCORR and MGC satisfy Theorem 2 directly, while HHG is also a statistic that is 0 if and only if independence (Heller et al., 2013).

(2) When $\mathbf{A}$ is symmetric and binary, the transformation from $\mathbf{A}$ to $\mathbf{L}$ is injective, i.e., two different $\mathbf{A}$ always produce two different $\mathbf{L}$. Then for each unique $\mathbf{L}$, the eigen-decomposition is always unique such that $\mathbf{L}$ to $\mathbf{U}^{t=1}$ is injective, provided that the dimension choice is made correct at q. □

(Corollary 2). From Proposition 1 and 2, $\mathbf{U}^{t=1}$ is asymptotically equivalent to the latent positions $\mathbf{W}$ up-to a bijection. Moreover, under RDPG, if two different adjacency matrices yield the same $\mathbf{U}^{t=1}$, they must asymptotically equal the same latent positions and asymptotically the same adjacency matrix (i.e., the difference in Frobenius norm converges to 0). Therefore injective holds asymptotically, and Theorem 3 applies. □

A.2 Additional Simulation

In order to investigate the performance of test statistics under the violation of non-positive semi-definite link function, i.e. under non-RDPG, we generate following stochastic block model (SBM) with two blocks for $i, j = 1, 2, \ldots, n = 100$:

$$
\begin{aligned}
&Z_i \overset{i.i.d.}{\sim} \mathbf{B}(0.5) \\
&\mathbf{A}(i,j) \mid Z_i, Z_j \sim Bernoulli\left((0.5-\epsilon)\mathbf{I}\left(|Z_i - Z_i| = 0\right) + 0.3\mathbf{I}\left(|Z_i - Z_j| \neq 0\right)\right) \\
&X_i \mid Z_i \sim \mathbf{B}\left(Z_i/3\right),
\end{aligned}
\tag{A.4}
$$

where Z_i represents block membership and nodal attribute of X_i depends on Z_i. Above Equation A.4 results non-positive semi-definite graph when $\epsilon > 0.2$, and beyond $\epsilon > 0.2$ increasing ϵ implies larger dependency.

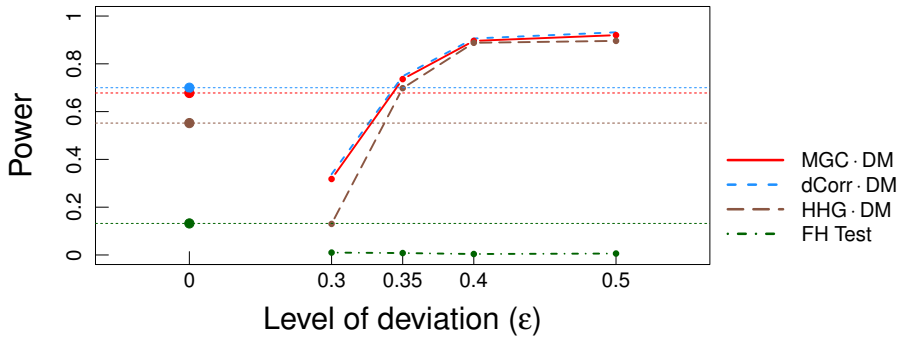


Figure A.1: Note that discrepancy in edge probability between the two blocks is both 0.2 when $\epsilon = 0$ and $\epsilon = 0.4$. Whereas MGC, DCORR, and HHG achieve higher power at $\epsilon = 0.4$ than $\epsilon = 0.0$, FH test does not work well under $\epsilon > 0.2$. Here testing power is empirically derived from $m = 500$ random replicates of which p-value is from $r = 500$ permutation samples.

Figure A.1 shows that distance-based methods, i.e., MGC, DCORR, and HHG, all preserve testing power under $\epsilon > 0.2$; while likelihood-based test of FH-test does not.

A.3 Random Dot Product Graph Simulations

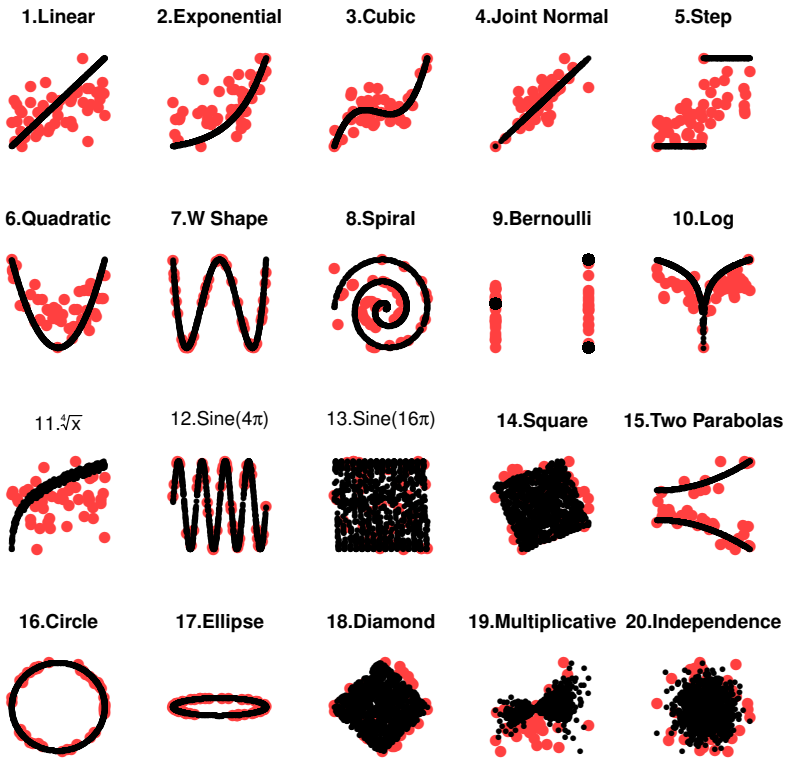


Figure A.2: Illustrations of randomly generated $n = 50$ points of $\{(W_i, X_i) : i = 1, 2, \ldots, 50\}$ (red dots) along with their population version without noise (black dots).

For the 20 simulations under RDPG, we describe the generating distribution $(\tilde{W}_i, \tilde{X}_i) \overset{i.i.d.}{\sim} F_{\tilde{W},\tilde{X}}$ under each scenario. Visualization for the sample observations of $\{(W_i, X_i) : i = 1, 2, \ldots, n = 50\}$ is shown in Figure A.2. Notation-wise, $\mathrm{N}(\mu, \sigma)$ denotes the normal distribution with mean μ and standard deviation σ, $\mathrm{U}[a, b]$ denotes the uniform distribution from a to b, $\mathrm{B}(p)$ denotes the Bernoulli

distribution with probability p, and ϵ_i denotes white noise.

1. Linear

$$\tilde{W}_i \sim \mathbf{U}[0,1],\ \epsilon_i \sim \mathbf{N}(0,0.5),$$
$$\tilde{X}_i = \tilde{W}_i + \epsilon_i.$$

2. Exponential

$$\tilde{W}_i \sim \mathbf{U}[0,3],\ \epsilon_i \sim \mathbf{N}(0,5),$$
$$\tilde{X}_i = \exp(\tilde{W}_i) + \epsilon_i.$$

3. Cubic

$$\tilde{W}_i \sim \mathbf{U}[0,1],\ \epsilon_i \sim \mathbf{N}(0,0.5),$$
$$\tilde{X}_i = 20(\tilde{W}_i - 0.5)^3 + 2(\tilde{W}_i - 0.5)^2 - (\tilde{W}_i - 0.5) + \epsilon_i.$$

4. Joint Normal

$$(\tilde{W}_i, \tilde{X}_i) \sim \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.7 & 0.5 \\ 0.5 & 0.7 \end{bmatrix}\right).$$

5. Step Function

$$\tilde{W}_i \sim \mathbf{U}[-1,1],\ \epsilon_i \sim \mathbf{N}(0,0.5),$$

$$\tilde{X}_i = \mathbf{I}(\tilde{W}_i > 0) + \epsilon_i.$$

6. Quadratic

$$\tilde{W}_i \sim \mathbf{U}[-1,1],\ \epsilon_i \sim \mathbf{N}(0,0.3),$$

$$\tilde{X}_i = \tilde{W}_i^2 + \epsilon_i.$$

7. W Shape

$$\tilde{W}_i \sim \mathbf{U}[-1,1]$$

$$\tilde{X}_i = 4(\tilde{W}_i^2 - 0.5)^2$$

8. Spiral

$$Z_i \sim \mathbf{U}[0,5],\ \epsilon_i \sim \mathbf{N}(0,0.1),$$

$$\tilde{W}_i = Z_i \cos(Z_i\pi),$$

$$\tilde{X}_i = Z_i \sin(Z_i\pi) + \epsilon_i.$$

9. Bernoulli

$$\tilde{W}_i \sim \mathbf{B}(0.5),\ \epsilon_i \sim \mathbf{N}(0,1),$$

$$\tilde{X}_i = (2\mathbf{B}(0.5) - 1)\tilde{W}_i + \epsilon_i.$$

10. Logarithm

$$\tilde{W}_i \sim \mathbf{U}[-1,1],\ \epsilon_i \sim \mathbf{N}(0,5),$$

$$\tilde{X}_i = 5\log_2(|\tilde{W}_i|) + \epsilon_i.$$

11. Fourth Root

$$\tilde{W}_i \sim \mathbf{U}[0,1],\ \epsilon_i \sim \mathbf{N}(0,0.5),$$

$$\tilde{X}_i = \left(|\tilde{W}_i + \epsilon_i|\right)^{1/4}.$$

12. Sine Period 4π

$$\tilde{W}_i \sim \mathbf{U}[-1,1],\ \epsilon_i \sim \mathbf{N}(0,0.01),$$

$$\tilde{X}_i = \sin(4\tilde{W}_i\pi) + \epsilon_i.$$

13. Sine Period 16π

$$\tilde{W}_i \sim \mathbf{U}[-1,1],\ \epsilon_i \sim \mathbf{N}(0,0.01),$$

$$\tilde{X}_i = \sin(16\tilde{W}_i\pi) + \epsilon_i.$$

14. Square

$$U_{i1} \sim \mathbf{U}[-1,1],\ u_{i2} \sim \mathbf{U}[-1,1],$$

$$\tilde{W}_i = U_{i1}\cos(-\pi/8) + U_{i2}\sin(-\pi/8),$$

$$\tilde{X}_i = -U_{i1}\sin(-\pi/8) + U_{i2}\cos(-\pi/8).$$

15.Two Parabolas

$$\tilde{Z}_i \sim \mathbf{B}(0.3),\ \epsilon_i \sim \mathbf{N}(0.5,0.3),$$

$$\tilde{W}_i \sim \mathbf{U}[0,1],$$

$$\tilde{X}_i = (\tilde{W}_i^2 + \epsilon_i)(\tilde{Z}_i - 0.5).$$

16. Circle

$$U_i \sim \mathbf{U}[-1,1],\ \epsilon_i \sim \mathbf{N}(0,0.05),$$

$$\tilde{W}_i = \cos(U_i\pi),$$

$$\tilde{X}_i = \sin(U_i\pi) + \epsilon_i.$$

17. Ellipse

$$U_i \sim \mathbf{U}[-1,1],$$

$$\tilde{W}_i = 5\cos(U_i\pi),$$

$$\tilde{X}_i = \sin(U_i\pi).$$

18. Diamond

$$U_{i1} \sim \mathbf{U}[-1,1],\ U_{i2} \sim \mathbf{U}[-1,1],$$

$$\tilde{W}_i = U_{i1}\cos(-\pi/4) + U_{i2}\sin(-\pi/4),$$

$$\tilde{X}_i = -U_{i1}\sin(-\pi/4) + U_{i2}\cos(-\pi/4).$$

APPENDIX A. SUPPLEMENTARY MATERIAL OF CHAPTER 4

19. Multiplicative Noise

$$\tilde{W}_i \sim \mathbf{N}(0.5, 1),\ \epsilon_i \sim \mathbf{N}(0.5, 1),$$

$$\tilde{X}_i = \tilde{W}_i \cdot \epsilon_i$$

20. Independence

$$\tilde{W}_i \sim \mathbf{N}(0, 1)$$

$$\tilde{X}_i \sim \mathbf{U}(0, 1)$$

Appendix B

Chain Graphs and Causal Inference in Social Network

This is a joint work in collaboration with Elizabeth Ogburn and Ilya Shpitser, and this Appendix presents a part of Ogburn et al. (2018b).

B.1 Graphs and Graphical Models

Graphical models use graphs–collections of vertices, representing random variables, and edges representing relations between pairs of vertices–to concisely represent conditional independences that hold among the random variables. At their most general, the graphical models we will consider in this paper are represented by mixed graphs containing directed ($\rightarrow$), and undirected

($-$) edges, such that at most one edge connects two vertices. In this section we review necessary concepts and terminology.

A sequence of non-repeating vertices $(V_1, \ldots, V_k)$ is called a *path* if for every $i = 1, \ldots, k-1$, V_i and V_{i+1} are connected by an edge. A path is *partially directed* if there exists an ordering of the vertices such that all directed edges in the path point towards the vertex with a larger index. A partially directed path is *directed* if it contains no undirected edges. A mixed graph is contains a *partially directed cycle* if it contains a partially directed path with a directed edge from the last to the first node in the path. A mixed graph with no partially directed cycles is called a *chain graph* (CG). A chain graph without undirected edges is called a *directed acyclic graph* (DAG), and a chain graph without directed edges is an *undirected graph* (UG).

If an edge $A \to B$ exists in a graph $\mathcal{G}$, A is a *parent* of B, and B is a *child* of A. If an edge $A - B$ exists in $\mathcal{G}$, then A is a *neighbor* of B (and vice versa). The sets of parents and children of A in $\mathcal{G}$ are denoted by $pa_{\mathcal{G}}(A)$ and $ch_{\mathcal{G}}(A)$, respectively. We define these sets on sets of vertices disjunctively, e.g. for a set of vertices $\mathbf{A}$, $pa_{\mathcal{G}}(\mathbf{A}) \equiv \bigcup_{A \in \mathbf{A}} pa_{\mathcal{G}}(A)$.

Consider an edge subgraph of a CG $\mathcal{G}$ that drops all directed edges and retains undirected edges. A connected component in such a subgraph is called a *block*. The set of blocks in a CG $\mathcal{G}$ will be denoted by $\mathcal{B}(\mathcal{G})$. This set partitions the set of vertices in $\mathcal{G}$. In an undirected graph $\mathcal{G}$, a *clique* is a maximal fully

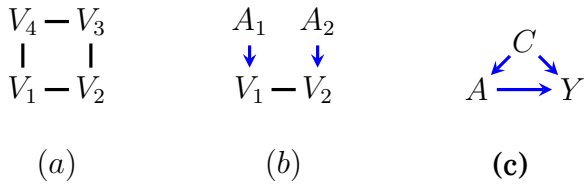


Figure B.1: (a) A simple undirected graph. (b) The simplest chain graph with an independence model not representable as either a DAG or an undirected graph. (c) A causal graph representing observed confounding of the treatment A and outcome Y by a set of covariates C.

connected set of vertices. Let the set of cliques in the edge subgraph be $\mathcal{C}(\mathcal{G})$. Note that unlike $\mathcal{B}(\mathcal{G})$, $\mathcal{C}(\mathcal{G})$ is not necessarily a partition of vertices in $\mathcal{G}$.

Graphical models encode conditional independences that hold in $p(\mathbf{V})$, the joint distribution of the random variables corresponding to the vertices of the graph. When a conditional independence holds, it tells us something about how $p(\mathbf{V})$ can be factorized, informing choices of models for $p(\mathbf{V})$. In the next sections, we describe results that translate the conditional independences encoded in a graphical model into a factorization of $p(\mathbf{V})$, and describe models for $p(\mathbf{V})$ that are consistent with the factorization. Any joint density that can be written according to that factorization will be consistent with the graphical model.

B.1.1 Directed acyclic graph models and causal inference

Given DAG $\mathcal{G}$, a DAG model is a set of distributions $p(\mathbf{V})$ that satisfy the following *Markov factorization*:

$$p(\mathbf{V} = \mathbf{v}) = \prod_{V \in \mathbf{V}} p(V \mid pa_{\mathcal{G}}(V)). \tag{B.1}$$

That is, the joint distribution of $\mathbf{V}$ is given by the product of the conditional distributions of each node given its parents. Each conditional distribution can be modeled directly to get parsimonious models for joint distributions over DAGs.

If a path in a DAG includes X, W and Z and if there are arrows from both X and Z into W, then W is a *collider* on the path. A path can be *unblocked*, meaning roughly that information can flow from one end to the other, or *blocked*, meaning roughly that the flow of information is interrupted at some point along the path. If all paths between two variables are blocked, then the variables are *d-separated*, and if two variables are d-separated then they are statistically independent. A path is blocked if there is a collider on the path such that neither the collider itself nor any of its descendants is conditioned on. An unblocked path can be blocked by conditioning on any noncollider along the path. Two variables are d-separated by a set of variables if conditioning on the variables in the set suffices to block all paths between them, and if two

variables are d-separated by a third variable or a set of variables then they are independent conditional on the third variable or set of variables (Pearl, 2000).

DAGs are powerful tools for causal inference using observational data and have gained widespread use in epidemiology, social sciences, and other fields, because they can be used to determine whether and how a counterfactual quantity can be identified from observed data. A primary object of interest in causal inference is a *counterfactual* or *potential outcome* $Y(a)$, which is a random variable representing the outcome, Y, that would have been observed if, possibly contrary to fact, an exposure or treatment A had been set to a. For a binary treatment, causal effects of A on Y can be expressed as contrasts between $E[Y(1)]$ and $E[Y(0)]$.

Inferences about counterfactuals are possible under assumptions linking the distribution $p(\mathbf{V})$, with $\{Y, A\} \subseteq \mathbf{V}$, representing factual data we observe, and the counterfactual distributions $p(Y(a))$ for any a in the support of A. The consistency assumption states that if the event $A = a$ is observed, then $Y(a) = Y$. In other words, the response Y does not distinguish between counterfactual assignment and factual occurrence of any value a. In the case of binary A, consistency entails that we observe one of the two counterfactual responses $Y_i(1), Y_i(0)$ for every unit i. The specific response we obtain corresponds to actually observed value of A_i (treatment for that unit). Consistency on its own is insufficient to make inferences about the ACE, as it gives us

only half of the relevant information. The assumption of *conditional ignorability* (Rosenbaum and Rubin, 1983), or *no unmeasured confounders*, states that $A \perp\!\!\!\perp \{Y(1), Y(0)\} \mid \mathbf{C}$, for a set of observed variables $\mathbf{C}$. Conditional ignorability is meant to represent the situation where A is not randomized and the treatment assignment process and the outcomes are confounded, but all sources of confounding are observed and contained in $\mathbf{C}$. Under this assumption, we have the following derivation:

$$p(Y(a)) = \sum_{\mathbf{C}} p(Y(a) \mid \mathbf{C}) p(\mathbf{C}) = \sum_{\mathbf{C}} p(Y(a) \mid A = a, \mathbf{C}) p(\mathbf{C}) = \sum_{\mathbf{C}} p(Y \mid a, \mathbf{C}) p(\mathbf{C}).$$

This is sometimes known as the *adjustment formula* or *backdoor formula*.

A DAG is called a *causal DAG* if it includes all common parents of any node in the graph. Whether conditional ignorability holds for a particular treatment-outcome relation can be read off of a causal DAG via the *backdoor criterion* (see Pearl (2000)): if all of the paths from the treatment to the outcome that begin with an arrow pointing *into* treatment can be blocked by observed covariates, then conditional ignorability holds. As a simple example, a setting where conditional ignorability hold are represented by a DAG in Figure B.1 (c). The directed arrows in such graphs can be interpreted informally to mean direct causation (see e.g. (Richardson and Robins, 2013) for a precise interpretation). In Figure B.1 (c), C acts as an observed common cause of A and Y, and

therefore any observed association between A and Y could be due to either the causal relationship of A and Y, represented by a directed edge between them, or to the non-causal dependence induced by C.

The *g-formula* (Robins, 1986) generalizes the adjustment formula to describe the relationship between the observed data distribution $p(\mathbf{V})$ and distributions of counterfactual random variables of the form $\{\mathbf{V} \setminus \mathbf{A}\}(\mathbf{a}) \equiv \{V(a)|V \in V\ A\}$. The intervention operation that sets a variable A to a can be viewed as replacing the distribution $p(A \mid pa_{\mathcal{G}}(A))$ by a deterministic distribution $p(A = a) = 1$, and all distributions $\{p(V \mid pa_{\mathcal{G}}(V)\}$ by distributions $\{p(V \mid pa_{\mathcal{G}}(V) \setminus \{A\}, a)\}$ for $V \in ch_{\mathcal{G}}(V)$. Generalizing this reasoning to a set of variables $\mathbf{A}$ being intervened on to attain a set of values a results in the g-formula:

$$p\left(\{\mathbf{V} \setminus \mathbf{A}\}(\mathbf{a}) = \mathbf{v}\right) = \prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V = v \mid pa_{\mathcal{G}}(V) \setminus \mathbf{A}, \mathbf{a}_{\mathbf{A} \cap pa_{\mathcal{G}}(V)}) \tag{B.2}$$

where $\mathbf{a}_{\mathbf{A} \cap pa_{\mathcal{G}}(V)}$ denotes the intervention values for the subset of $\mathbf{A}$ that intersects with the parents of V.

Typically in causal inference applications, we are interested in the counterfactual response of a single outcome variable $Y \in \mathbf{V} \setminus \mathbf{A}$ to an intervention that sets $\mathbf{A}$ to $\mathbf{a}$, which can easily be obtained from the g-formula, especially for a low dimensional outcome Y. But in settings with complex networks of outcomes, e.g. representing systems of agents interacting with one another, statistical

inference about the g-formula is impractical or impossible, and other tools are needed.

B.1.2 Undirected graph and chain graph models

Given an undirected graph $\mathcal{G}$, an undirected graphical model is a set of distributions $p(\mathbf{V})$ that satisfy the *global Markov property* (Lauritzen, 1996): each node is independent of its non-neighbors conditional on its neighbors. That gives the following *clique factorization*:

$$p(\mathbf{V}) = \frac{1}{Z} \prod_{\mathbf{C} \in \mathcal{C}(\mathcal{G})} \phi_{\mathbf{C}}(\mathbf{C}). \tag{B.3}$$

Any undirected graphical models can be written as a log-linear model, with a term for each clique in the factorization:

$$p(\mathbf{V} = \mathbf{v}) = \frac{1}{Z} \exp \left\{ \sum_{\mathbf{C} \in \mathcal{C}(\mathcal{G})} \log \phi_{\mathbf{C}}(\mathbf{v}_{\mathbf{C}}) \right\}, \tag{B.4}$$

where Z is a normalizing constant. This form implies conditional independence constraints on $p(\mathbf{V})$ via the global Markov property on $\mathcal{G}$.

For example, the factorization for the grid graph in Figure B.1 (a) can be

re-expressed as

$$\begin{aligned} p(\mathbf{V} = (v_1, v_2, v_3, v_4)) &= \frac{1}{Z}\phi_{1,2}(v_1, v_2)\phi_{2,3}(v_2, v_3)\phi_{3,4}(v_3, v_4)\phi_{1,4}(v_1, v_4) \\ &= \frac{1}{Z}\exp\left\{\log\phi_{1,2}(v_1, v_2) + \log\phi_{2,3}(v_2, v_3) + \log\phi_{3,4}(v_3, v_4) + \log\phi_{1,4}(v_1, v_4)\right\} \\ &= \frac{1}{Z}\exp\{h_1v_1 + h_2v_2 + h_3v_3 + h_4v_4 \\ &\quad + k_{1,2}v_1v_2 + k_{2,3}v_2v_3 + k_{3,4}v_3v_4 + k_{1,4}v_1v_4\}, \end{aligned}$$

where without lack of generality we can assign h_iv_i to any $\log\phi_{ij}$. Conditional independence constraints $V_1 \perp\!\!\!\perp V_3 \mid V_2, V_4$ and $V_2 \perp\!\!\!\perp V_4 \mid V_1, V_3$ hold in any $p(v_1, v_2, v_3, v_4)$ with the above factorization.

Chain graphs allow both directed and undirected edges, and can be used to define hybrid graphical models combining features of both undirected graphs and DAGs (Lauritzen, 1996). Given a chain graph $\mathcal{G}$ with a vertex set $\mathbf{V}$, we say a distribution $p(\mathbf{V})$ is in the chain graph model of $\mathcal{G}$, if

$$p(\mathbf{V} = \mathbf{v}) = \prod_{\mathbf{B}\in\mathcal{B}(\mathcal{G})} p(\mathbf{B} \mid pa_{\mathcal{G}}(\mathbf{B})), \tag{B.5}$$

where each factor $p(\mathbf{B} \mid pa_{\mathcal{G}}(\mathbf{B}))$ further factorizes as

$$p(\mathbf{B} \mid pa_{\mathcal{G}}(\mathbf{B})) = \frac{1}{Z(pa_{\mathcal{G}}(\mathbf{B}))}\left(\prod_{\mathbf{C}\in\mathcal{C}((\mathcal{G}_{fa_{\mathcal{G}}(\mathbf{B})})^a), \mathbf{C}\not\subseteq pa_{\mathcal{G}}(\mathbf{B})} \phi_{\mathbf{C}}(\mathbf{v}_{\mathbf{C}})\right), \tag{B.6}$$

$Z(\mathbf{v}_{pa_{\mathcal{G}}(\mathbf{B})})$ is a mapping from values of $pa_{\mathcal{G}}(\mathbf{B})$ to appropriate normalizing con-

stants, $(\mathcal{G}_{fa_{\mathcal{G}}(\mathbf{B})})^a$ is an undirected graph consisting of vertices in $fa_{\mathcal{G}}(\mathbf{B}) \equiv \mathbf{B} \cup pa_{\mathcal{G}}(\mathbf{B})$, and an undirected edge between any pair in $fa_{\mathcal{G}}(\mathbf{B})$ adjacent in $\mathcal{G}$ or any pair in $pa_{\mathcal{G}}(\mathbf{B})$.

The chain graph factorization can be viewed as a two-level factorization. The outer factorization (B.5) resembles the Markov factorization for DAG models (B.1) (Pearl, 1988), while the inner factorization (B.6) for each outer factor $p(\mathbf{B} \mid pa_{\mathcal{G}}(\mathbf{B}))$ resembles the undirected factorization (B.3). The chain graph factorization of $p(\mathbf{V})$ induces a set of conditional independences on $p(\mathbf{V})$ via a global Markov property just as was the case for undirected models, although this property is more involved to define (see Lauritzen (1996) for details).

For example, the chain graph in Figure B.1 (b) has the factorization

$$p(v_1, v_2, a_1, a_2) = \left(\frac{1}{Z(a_1, a_2)} \exp\left\{ \phi_{v_1,v_2}(v_1, v_2) \phi_{v_1,a_1}(v_1, a_1) \phi_{v_2,a_2}(v_2, a_2) \right\} \right) p(a_1) p(a_2).$$

and implies that the conditional independences $V_1 \perp\!\!\!\perp A_2 \mid A_1, V_2$ and $V_2 \perp\!\!\!\perp A_1 \mid A_2, V_1$ hold in $p(v_1, v_2, a_1, a_2)$.

Undirected and chain graph models have a deceptively intuitive appeal for modeling social network data. At first glance the global Markov property seems like a reasonable way to impose statistical structure on the ties in a social network: it implies that each node is "screened off" from its non-neighbors given its neighbors, which sounds consistent with a process of influence where each

node can only affect its neighbors and any longer range dependence is mediated by paths from one node, through its neighbors, to the broader network. However, undirected edges are not consistent with the causal influence of one individual on another. Indeed, Lauritzen and Richardson (2002) argue that many seemingly intuitive uses for the undirected edges in undirected and chain graphs are in fact misguided. Undirected edges have been used to represent symmetric associations, non-causal associations, simultaneous responses, processes with feedback, ignorance of the direction of arrow between two nodes, and causal relations that can change directions, but all of these are inconsistent with chain graph models. Importantly, they argue that there are no chain graph models consistent with most DAG models: these two classes of models represent largely non-overlapping classes of joint distributions. Even projecting a DAG model onto a subset of variables in the model cannot generally result in a chain graph model. Instead, the undirected edges in chain graphs represent certain kinds of equilibria, some examples of which are described in Lauritzen and Richardson (2002).

B.1.3 Graphical models for social interactions

Causal DAG models, or the mathematically equivalent causal structural equation models, are assumed either implicitly or explicitly in almost all existing methods for learning about social interactions, interference, and contagion

from observational data. DAGs and causal structural equation models correspond to a mechanistic view of the (macroscopic) world, which is espoused by most researchers across many disciplines. In particular, almost all approaches to learning about causal effects from data assume are based on this mechanistic view of the world. The impact that one individual has on another is a causal effect, and therefore most of the literature on social influence makes use of causal ideas, terminology, and methods (though sometimes not overtly).

Ogburn et al. (2014) is an overview of the use of DAGs to represent interference and contagion. New methods for learning about spillover and contagion effects from social network data similarly rely on assumptions that are consistent with DAG models but not with CG models, explicitly in the case of methods for observational data proposed by van der Laan (2014) and Ogburn et al. (2017) and implicitly in many of the methods based on randomized experiments (e.g. Aronow and Samii, 2012; Athey et al., 2016; Bowers et al., 2013; Choi, 2014; Eckles et al., 2014; Forastiere et al., 2016; Graham et al., 2010; Hong and Raudenbush, 2006, 2008; Hudgens and Halloran, 2008; Jagadeesan et al., 2017; Liu and Hudgens, 2014; Liu et al., 2016; Rosenbaum, 2007; Rubin, 1990a; Sobel, 2006; Tchetgen Tchetgen and VanderWeele, 2012; VanderWeele, 2010). However, as we will show in the next section and as has been acknowledged by some of the aforementioned researchers, DAGs in these settings can quickly become cumbersome.

B.2 Chain Graph Approximation

In contrast to classical causal inference, where treatments and outcomes are independent across subjects, we are interested in representing and reasoning about situations where outcomes are complicated and may represent dependent processes across individuals connected by social ties. Consider a social network of n individuals, or nodes. Node i is associated with a treatment or exposure, A_i, an outcome Y_i, and possibly covariates. For example, Y could represent opinions and A advertising campaigns; Y could represent behavior and A encouragement interventions, or Y could represent an infectious disease and A vaccination. We represent a set of outcomes on individuals in a social network by vertices connected by undirected edges. In addition, we want to represent causal influences of interventions on these outcomes, and variables that may serve as confounders for such influences in observed data. Edges involving these variables will be directed, representing causality. When individual's beliefs or opinions undergo phase transitions to orderly states, e.g. when there is external pressure to reach a unanimous consensus, or when it can be argued that the distribution of individual's behaviors, beliefs, opinions, or other outcomes attains an equilibrium across network ties, then a chain graph may be the correct model for the joint distributions of outcomes across a network and interventions on those outcomes. For example, in the Supreme

APPENDIX B. CHAIN GRAPHS AND CAUSAL INFERENCE IN SOCIAL NETWORK

Court data that we analyze below, outcomes represent decisions made under time constraints and with pressure for the nine justices to reach a unanimous decision; these may indeed be in equilibrium. More common in the existing literature are settings in which DAG models would be the most appropriate but are not tractable given reasonable constraints on data collection.

We make the routine assumption that interference can occur only directly between two individuals who share a tie in the underlying social network. That is, any effect of an ego's treatment on a non-alter's outcome must be mediated by mutual connections. Figure B.2 (a) depicts one of the simplest such settings: the network is comprised of only three individuals; individuals 1 and 2 share a tie and 2 and 3 share a tie; each individual's outcome is affected by her own treatment, her own past outcomes, and her social contacts' past outcomes. In order for this DAG to be valid, the units of time captured must be small enough that any influence passing from 1 to 3 through 2 cannot occur in fewer than 2 time steps (Ogburn et al., 2014). This will be the case if influence can only occur during discrete interactions such as in-person or online encounters, and the unit of time is chosen to be the minimum time between encounters. This DAG model encodes several conditional independences, and if we are able to observe the outcome for all agents at all time steps, inference under these models may be possible (Ogburn et al., 2017).

However, in most practical applications, with the exception of online social

networks, it is only be possible to observe the outcome at one or a few time points. If data are generated according to the DAG in Figure B.2 (a) but the outcome is observed at only one time point (at which the outcome is not in a chain graph equilibrium), then the resulting model is represented by a mixed graph representing the *latent projection* of all of the variables in Figure B.2 (a) onto the subset of those variables that are actually observed, with *bidirected* edges representing the presence of one or more hidden common causes (Spirtes and Verma, 1992). A general construction algorithm for these latent projection mixed graphs is given by Pearl (2009), and the result for Figure B.2 (a) is shown in Figure B.2 (b).

Collecting or accessing the detailed temporal data required to use the models like Figure B.2 (a) is often impractical or impossible, but the saturated model for the marginal in Figure B.2 (b) quickly becomes unwieldy, as the number of parameters required to estimate and use the model grows exponentially with the number of nodes: the latent projection graph will generally not be sparse, even if the underlying social network governing opinion formation is. To see this, note that after a single time step, an individual only influences neighboring individuals, but after two time steps, also neighbors of neighbors. In the three-person network represented by Figure B.2, this is enough to render the latent projection of Fig B.2 (b) fully saturated, with no conditional independences. After many time steps, the individual's influence would have time to

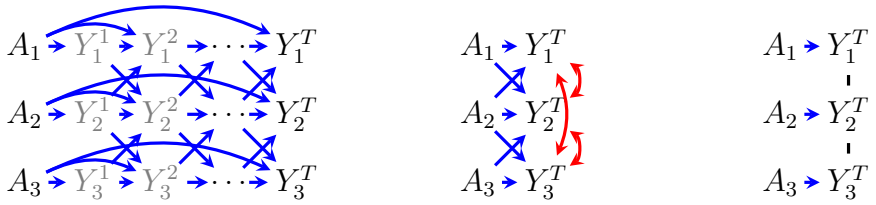


Figure B.2: (a) Causal DAG representing opinion formation among peers. A_i represents interventions meant to influence subject i, Y_i^k is the ith subject's opinion at time k. (b) A latent projection of the model in (a) onto variables $A_1, A_2, A_3, Y_1^T, Y_2^T, Y_3^T$, representing the distribution of opinion in (a) at time T, before equilibrium is reached. The red bidirected arrows represent the fact that the outcomes at intermediate time points are unmeasured common causes of the observed outcomes. (c) A chain graph model that approximates the distribution of opinion in (a) at time T under certain data generating processes.

reach most of the social network. This implies that any two outcomes at time t, for a large enough t, will be related via a chain of hidden common causes, even if the corresponding individuals are far from each other in the social network. To represent these chains of hidden common causes, the latent projection graph would contain a clique of bidirected edges encompassing opinions of everyone in the network. The significance of a non-sparse latent projection is that the corresponding statistical model is has exponentially many parameters, even if all variables are binary. These limitations are reflected in the literature, which rarely includes applications to real data.

Unlike the model in Figure B.2 (b), the chain graph model represented by Figure B.2 (c) is not saturated. In certain cases, a chain graph model that is as sparse as the underlying social network may serve as a good approximation of the intractable latent projection model.

APPENDIX B. CHAIN GRAPHS AND CAUSAL INFERENCE IN SOCIAL NETWORK

Consider chain graph models like the one in corresponding to Figure B.2 (c), but with arbitrary undirected components corresponding to outcomes observed on social network nodes. These models imply that each node's outcome is independent of its non-neighbors' outcomes conditional on its neighbors' outcomes and on any treatments or covariates with arrows pointing into the node. For chain graphs like Figure B.2 (c), with a single treatment for each node, the conditional independences implied by the graph are of the form $Y_i^T \perp\!\!\!\perp A_j, Y_j^t \mid A_i, \{Y_l^T, \forall l \text{ adjacent to } i\}$. These conditional independences fail to hold in the corresponding DAG models due to two different types of paths, depicted in red in Figure B.3.

Paths like the one in Figure B.3 (a) represent the fact that the past outcomes of mutual connections affect both Y_i^T and Y_j^T; this is just one of many such paths. All of these paths can be blocked by conditioning on $\{Y_l^t, \text{ for all } l \text{ adjacent to } i \text{ and for } 1 \leq t \leq T-1\}$ (Ogburn and VanderWeele, 2017). If the outcome evolves slowly over time, Y_l^t and Y_l^T will be highly correlated and conditioning on $\{Y_l^T, \text{ for all } l \text{ adjacent to } i\}$ will mostly block these paths. We expect the paths through Y_l^t to be weaker for smaller t than for t close to T. If paths through Y_l^t are weaker for earlier times t, then the relationship between Y_l^t and Y_l^T can also weaken for decreasing t – as long as it remains strong enough to allow conditioning on Y_l^T to approximately block paths through Y_l^t.

However, conditioning on $\left\{Y_l^T, \text{ for all } l \text{ adjacent to } i\right\}$ opens paths through colliders, like the one depicted in Figure B.3 (b). M-shaped collider paths like these are known to induce weak dependence in general (Greenland, 2003), and the magnitude can be bounded more precisely using knowledge of the partial correlation structure of the variables along the path (Chaudhuri and Richardson, 2002). Informally, if the dependence of Y_l^T on Y_l^{T-1} is stronger than that of Y_l^T on Y_i^{T-1} and Y_j^{T-1}, as it will be if the outcome evolves slowly over time, then the dependence induced by paths through colliders may be negligible.

Although chain graph models exist in which the relationships along undirected edges are not symmetric, we found in simulations that DAGs with symmetric relationships for connected pairs of individuals were better approximated by chain graphs. It might be reasonable to assume this kind of symmetry if, for example, the outcome is a behavior or belief and the subjects are peers with no imbalance of power or influence, or if the outcome is an infectious disease and the subjects have similar underlying health and susceptibility statuses.

To demonstrate how the data generated from DAG model can be successfully approximated by a chain graph, we simulated ten random nine-node graphs with edge probability $p = 0.3$ with nine agents. For each random graph, we generated outcomes for each node according to a DAG model independently 1000 times (See Equation B.7). For all nonadjacent pairs (i, j), we tested (1)

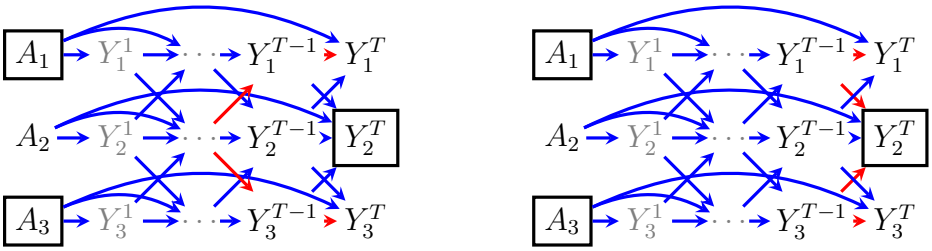


Figure B.3: Paths that connect Y_1^T and Y_3^T even when conditioning on Y_2^T and A_1 and/or A_3. Boxes indicate variables that are conditioned on.

the null hypothesis of marginal independence $Y_i^t \perp\!\!\!\perp Y_j^t$, (2) the null hypothesis of conditional independence $Y_i^t \perp\!\!\!\perp Y_j^t \mid \{Y_l^t, \forall l \text{ adjacent to } i\}$, and (3) the null hypothesis of conditional independence $Y_i^t \perp\!\!\!\perp Y_j^t \mid A_i, \{Y_l^t, \forall l \text{ adjacent to } i\}$.

$$A_i \overset{i.i.d.}{\sim} 2\mathcal{B}(0.5) - 1, \ t = 1, 2, \ldots, 9$$

$$Y_i^1 \overset{i.i.d.}{\sim} 2\mathcal{B}(0.5) - 1, \ t = 1, 2, \ldots, 9$$

For $t = 2, 3, \ldots, 100$:

$$Y_i^t \sim \begin{cases} \mathcal{B}\left(h\big(-0.5 + 5Y_i^{t-1} + 0.8A_i + 0.5 \sum_{j \in N(i)\setminus\{i\}} Y_j^{t-1} + \mathcal{N}(0, 0.1)\big)\right) & i = 1, \ldots, 4 \\ \mathcal{B}\left(h\big(0.0 + 5Y_i^{t-1} + 0.8A_i + 0.5 \sum_{j \in N(i)\setminus\{i\}} Y_j^{t-1} + \mathcal{N}(0, 0.1)\big)\right) & i = 5 \\ \mathcal{B}\left(h\big(0.5 + 5Y_i^{t-1} + 0.8A_i + 0.5 \sum_{j \in N(i)\setminus\{i\}} Y_j^{t-1} + \mathcal{N}(0, 0.1)\big)\right) & i = 6, \ldots, 9, \end{cases}$$

$$Y_i^t \leftarrow 2Y_i^t - 1,$$

(B.7)

where $\mathcal{B}(p)$ denotes Bernoulli distribution with probability p; $\mathcal{N}(\mu, \sigma)$ denotes

APPENDIX B. CHAIN GRAPHS AND CAUSAL INFERENCE IN SOCIAL NETWORK

Normal distribution with mean μ and standard deviation σ; $N(v)$ means a set of adjacent nodes of v. Conditional and marginal independence test results applied for all non-adjacent pairs are presented in Figure B.4.

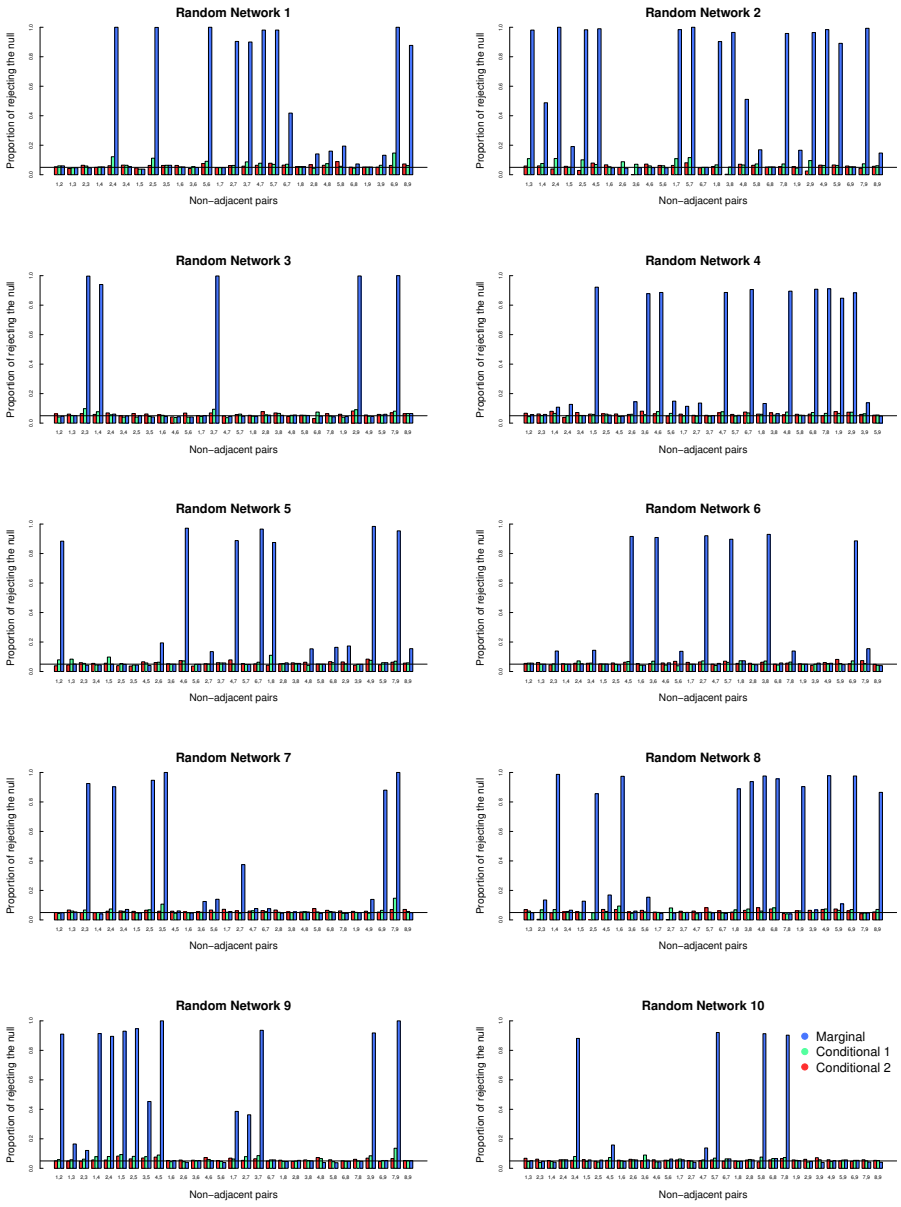


Figure B.4: Each bar plot shows the proportion of rejecting the null of marginal independence (blue), conditional independence 1 (green), conditional independence 2 (red) applied for all non-adjacent pairs from each random graph. Here marginal independence means $Y_i^t \perp Y_j^t$; conditional independence 1 means $Y_i^t \perp Y_j^t \big| Y_{N(i)\setminus\{i\}}^t$; and conditional independence 2 means $Y_i^t \perp Y_j^t \big| Y_{N(i)\setminus\{i\}}^t, A_i$. Marginally dependent non-adjacent pairs Y_i^t and Y_j^t are conditionally independent when conditioned on A_i as well as $Y_{N(i)\setminus\{i\}}^t$, maintaining nominal 5% of rejection rate (horizontal line). (Zero proportion of conditional independence tests in network 8 is due to large adjacent peers conditioned on compared to the sample size $n = 1000$.)

We found that the conditional independence nulls were rejected at close to the nominal rate of 5% expected under the null. In contrast, the marginal independence null was rejected more frequently, suggesting that conditioning on neighbors' outcomes may recover approximate independence under at least some data generating processes, and that the chain graph model may in those cases be a reasonable parsimonious approximation to the true underlying conditional independences.

B.3 Collective Decision Making in Supreme Court

The U.S. Supreme Court is comprised of nine justices, one of whom is the Chief Justice, tasked with presiding over oral arguments, serving as the spokesperson for the court, and other administrative roles. After a case is heard by the Supreme Court, the justices discuss and decide the case over a period of several weeks or months. The final outcome is decided by majority vote; the majority and, when the decision is not unanimous, the minority write opinions justifying their decisions. The oral and written arguments presented to the court and the judicial opinions are public resources; however, we have no access to the debates and discussions that lead the justices to their decisions. This precludes the use of a DAG model for the evolution of individuals' opinions over time, but

is amenable to a chain graph model with Y_i defined as Justice i's final opinion.

Data on all Supreme Court decisions since 1946, along with rich information on the nature of the cases and the opinions, is maintained by Washington University Law School's Supreme Court Database (`http://scdb.wustl.edu/data.php`). We used the subset of these data corresponding to the Second Rehnquist Court, a period of ten years (1994-2004) during which the same nine justices served together: William Rehnquist (Chief Justice), John Paul Stevens, Sandra Day O'Connor, Antonin Scalia, Anthony Kennedy, David Souter, Clarence Thomas, Ruth Bader Ginsburg, and Stephen Breyer. Over these ten years the court decided 893 cases.

The Supreme Court Database has classified each case into one of 14 issue areas, such as criminal procedure and civil rights. We examined the effect of issue area on conservative vs. liberal opinions. For each case, each justice has an outcome, Y, which is an indicator of a liberal opinion. The ruling in the case is liberal if at least 5 of the justices form liberal opinions and conservative otherwise. During the Rehnquist court, 56$\%$ of the decisions were conservative. Clarence Thomas was the most conservative justice, signing the conservative opinion in 72$\%$ of cases, while Ruth Bader Ginsburg was the most liberal, signing the liberal opinion in 60$\%$ of cases. However, we found that issue area had a strong effect on both individual outcomes and on overall court decisions, which is consistent with literature on the effect of issue areas on the ideology of each

justice or on the final decision of Supreme Court (Tate, 1981; Lu and Wang, 2011).

Issue area	Criminal procedure	Civil rights	First amendment	Due process	Privacy	Attorneys
|case|	231	161	59	43	21	5
Issue area	Unions	Economic Activity	Judicial Power	Federalism	Federal taxation	Total
|case|	18	145	133	57	20	893

Table B.1: The number of cases decided during 1994-2004. There is no case about Interstate relations, Miscellaneous, nor Private action.

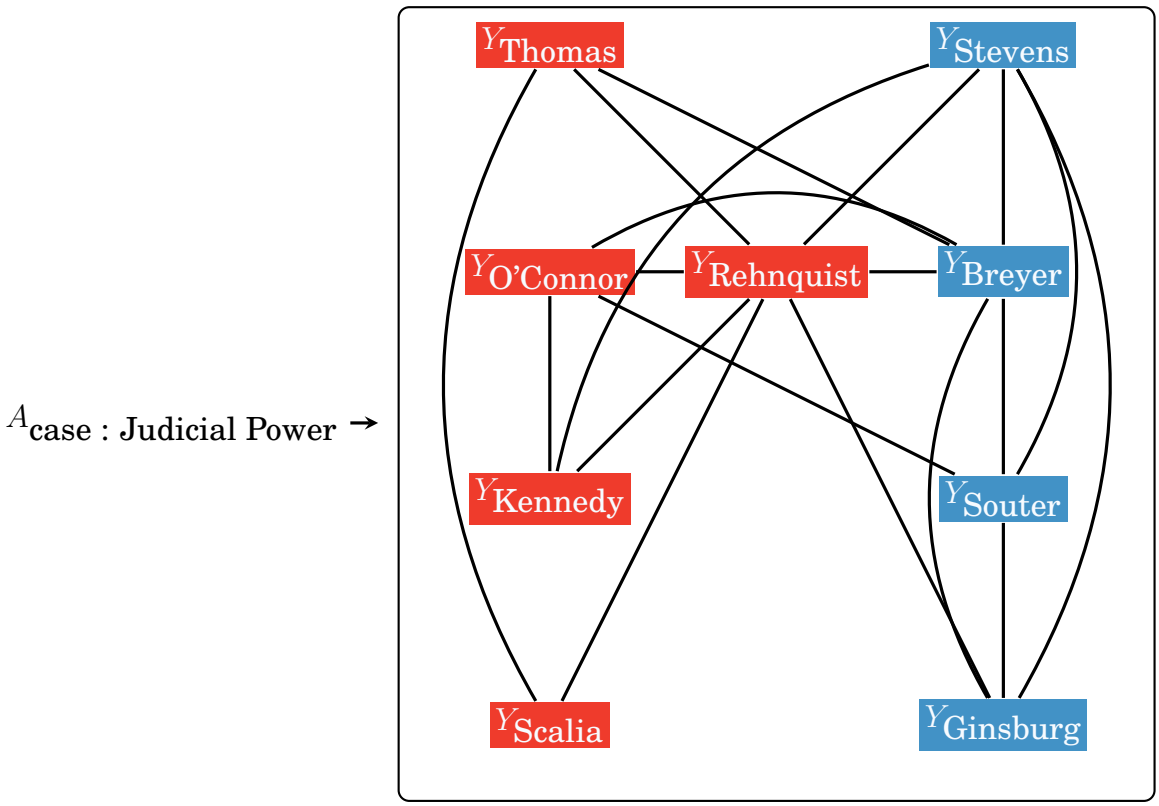


Figure B.5: The underlying network between nine justices assuming the model where the intervention of A indicates the case is about judicial power. The color of each node indicates well-known political orientation or political party affiliation – red indicates conservative or from republican party, and blue indicates liberal or from democratic party, but of course, we do not know the it really is. The undirected edge between the justices implies the existence of some interactions or feedback in decision making procedures learned through network structure learning procedure.

B.3.1 Causal inference on collective decisions

We will separately consider the effects of indicators of (i) criminal procedure, (ii) civil rights, (iii) economic activity, and (iv) judicial power on conservative vs liberal opinions. Although there is strong evidence (including self-report by the justices) that the Court works hard to come to unanimous decisions, 5-to-4 decisions are frequent (Sunstein, 2014; Riggs, 1992). There is also considerable academic interest in each justice's personal orientation (Songer and Lindquist, 1996; Tate, 1981). A chain graph model can answer causal questions such as: do any of the issue areas cause a significantly greater probability of unanimous decisions relative to the other areas?

First, we fit a log-linear model with all pairwise interaction terms in order to estimate the social network by which justices influence one another, with undirected edges between justices given by the magnitude of their interaction coefficient. (This ad-hoc method performed as well as established structure-learning algorithms in simulations and was easier to implement.) We used this estimated network as the undirected component of a chain graph and added a single treatment variable, i.e. issue area, that jointly affects each justice's outcome. The resulting chain graph for the judicial power issue area is displayed in Figure B.5. Informally, there seems to be a liberal (blue) clique and two conservative (red) cliques: a more moderate one comprised of O'Connor and Kennedy, and a more conservative one comprised of Scalia and Thomas–with

APPENDIX B. CHAIN GRAPHS AND CAUSAL INFERENCE IN SOCIAL NETWORK

Chief Justice Rehnquist serving as a hub with connections to almost every other justice (with the exception of Souter). We found that justices interact with one another not only based on their shared liberal or conservative leanings, but also across that divide. For example, Justices Stevens and Kennedy are known to have had different judicial philosophies and views, but there is anecdotal evidence that they often influenced one another's votes[1]. The tie between Breyer and O'Connor could be explained by their social connections[2] or their shared views on judicial independence[3]. Thomas and Breyer sat next to each other on the bench and were thought to have developed a close working relationship as a result[4].

Separately for each of the four issue areas, we estimated the parameters of the following chain graph model, based on the graph in Figure B.5:

$$p\big(\mathbf{Y} = (y_1, y_2, \ldots, y_9)|A = a\big) = \frac{1}{Z} \exp\left\{\sum_{i=1}^{9} h_i y_i + \sum_{i,j=1,e_{ij}=1}^{9} k_{ij} y_i y_j + \sum_{i=1}^{9} \gamma_i a y_i\right\}, \tag{B.8}$$

where $e_{ij} = 1$ implies justice i and j share an undirected edge in the chain graph. The parameter h_i represents the conservative or liberal leaning of Justice i, with a positive h_i indicating bias towards liberal opinions, and the inter-

[1] http://www.nytimes.com/2007/09/23/magazine/23stevens-t.html

[2] http://blogs.findlaw.com/supreme_court/2017/03/supreme-court-shutters-justice-oconnors-workout-class.html

[3] http://www.pbs.org/newshour/bb/law-july-dec06-independence_09-26/

[4] http://www.abajournal.com/news/article/breyer_sometimes_poses_questions_for_thomas_during_oral_arguments/

action parameter k_{ij} captures the tendency of Justices i and j to agree, with a positive k_{ij} indicating tendency to agree while a negative k_{ij} indicates tendency to disagree. The parameter γ_i is related to the causal effect of issue area a on Justice i's opinions, with positive γ_i indicating tendency toward liberal opinions above and beyond what can be explained by the Justice's independent leaning or by the interactions with other Justices. In principle three-way interactions could be added to the model to capture tendencies of groups of three justices to agree or disagree beyond what the pairwise interactions explain, but we did not have enough data to reliably estimate these additional parameters. We bootstrapped the standard errors in order to calculate 95% confidence intervals, with $nb = 500$ bootstrap samples for each model.

Table B.2 displays the main effects for each justice across four issue areas. As expected, Rehnquist, O'Connor, and Thomas tended towards opinions that were more conservative across all issue areas, while Stevens, Souter, and Ginsburg tended towards more liberal opinions. The direction of the main effect for Justices Scalia, Kennedy, and Breyer depends on the issue area. In Figure B.6 the shade of the node reflects the estimated main effect and the type and width of the edges reflects the magnitude and sign of the estimated interaction for $A = \mathbf{I}$(judicial power). The dotted edge connecting Rehnquist and Stevens represents the only negative interaction. Interestingly, the Rehnquist/Stevens interaction term is negative (and statistically significant) across all four issue

areas. This is corroborated by anecdotal evidence, as Stevens was reputed to be the most likely to disagree with the other justices[5] (Sirovich, 2003).

Issue	WHRehnquist	JPStevens	SDOConnor	AScalia	AMKennedy
Criminal procedure	-0.29 [-0.54 ,-0.12]	0.48 [0.35 ,0.63]	-0.33 [-0.50 ,-0.16]	-0.00 [-0.16 ,0.17]	-0.12 [-0.28 ,0.04]
Civil rights	-0.12 [-0.35 ,0.07]	0.51 [0.37 ,0.68]	-0.27 [-0.46 ,-0.10]	-0.02 [-0.19 ,0.15]	-0.21 [-0.38 ,-0.03]
Economic activity	-0.16 [-0.37 ,0.02]	0.33 [0.20 ,0.47]	-0.28 [-0.45 ,-0.12]	-0.00 [-0.17 ,0.18]	-0.05 [-0.20 ,0.11]
Judicial power	-0.20 [-0.46 ,0.02]	0.26 [0.13 ,0.40]	-0.23 [-0.43 ,-0.04]	-0.13 [-0.32 ,0.08]	-0.02 [-0.19 ,0.15]
Issue	DHSouter	CThomas	RBGinsburg	SGBreyer	
Criminal procedure	0.18 [0.00 ,0.39]	-0.42 [-0.61 ,-0.25]	0.30 [0.09 ,0.50]	0.03 [-0.15 ,0.23]	
Civil rights	0.27 [0.08 ,0.45]	-0.55 [-0.76 ,-0.37]	0.08 [-0.13 ,0.28]	0.15 [-0.03 ,0.35]	
Economic activity	0.07 [-0.08 ,0.24]	-0.29 [-0.46 ,-0.13]	0.31 [0.09 ,0.49]	0.01 [-0.18 ,0.19]	
Judicial power	0.20 [-0.00 ,0.40]	-0.28 [-0.49 ,-0.09]	0.14 [-0.09 ,0.33]	0.04 [-0.15 ,0.25]	

Table B.2: Coefficients and their 95% confidence intervals corresponding to personal orientation $\{k_i : i = 1, 2, \ldots, 9\}$ in model B.8.

[5] http://www.nytimes.com/2007/09/23/magazine/23stevens-t.html?mcubz=0

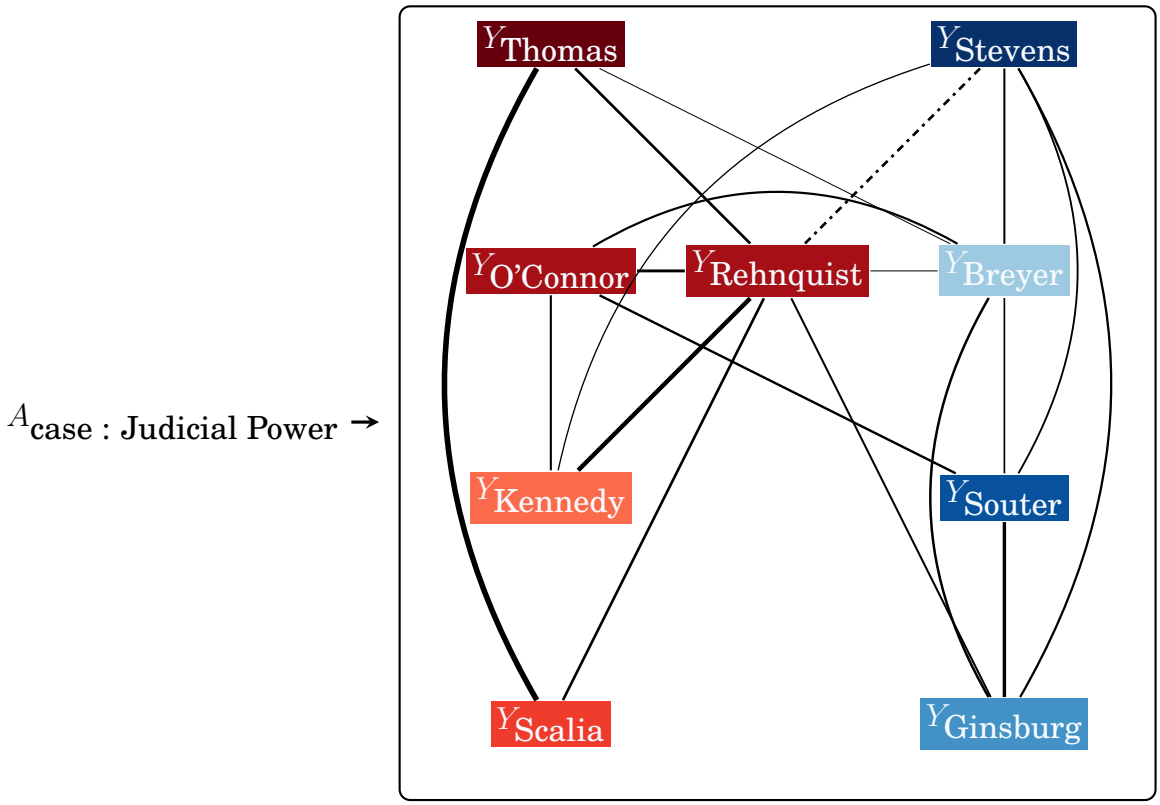


Figure B.6: The color of each node is shaded according to the estimated coefficients for the main effect in Table B.2. The darker the red node, the more conservative the corresponding justice's vote tends to be even with influence of others considered; similarly, the darker the blue node, the more liberal the justice's votes are. The width of the edge between justice i and j is weighted proportional to the absolute value of coefficients k_{ij}, and the edge between justice Rehnquist and justice Stevens is dashed due to the negative value of the coefficient.

Using the model given in Equation B.8, we estimated the causal effects of issue area on the majority-based decisions of the nine justices. We found that judicial power resulted in the highest probability of unanimous decisions, with those decisions more likely than baseline to be conservative (Table B.6). Economic activity resulted in a higher probability of liberal and unanimous

decisions than baseline (Table B.5). Criminal procedures and civil rights both increased the probability of 4 (liberal)-to-5 (conservative) decisions (Table B.3 and Table B.4).

Criminal procedure

$\|nl\|$	0	1	2	3	4
a=1	0.19 (0.02)	0.11 (0.01)	0.08 (0.01)	0.09 (0.01)	0.17 (0.02)
a=0	0.22 (0.02)	0.10 (0.01)	0.06 (0.00)	0.06 (0.00)	0.10 (0.01)
Causal effect	-0.03 (0.02)	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)	0.06 (0.02)
$\|nl\|$	5	6	7	8	9
a=1	0.10 (0.01)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)	0.11 (0.02)
a=0	0.07 (0.01)	0.06 (0.01)	0.07 (0.01)	0.06 (0.00)	0.18 (0.01)
Causal effect	0.02 (0.01)	-0.00 (0.01)	-0.03 (0.01)	-0.02 (0.01)	-0.07 (0.02)

Table B.3: The estimated probability of having unanimity (when $|nl| = 0, 9$) or having dissension when the case is about criminal procedure ($a = 1$) or others ($a = 0$) according to the number of liberal-side vote $|nl|$. The probability of unanimity towards liberal opinions decreases in this case.

Civil rights

$\|nl\|$	0	1	2	3	4
a=1	0.18 (0.02)	0.11 (0.02)	0.07 (0.01)	0.07 (0.01)	0.16 (0.02)
a=0	0.22 (0.01)	0.10 (0.01)	0.07 (0.00)	0.07 (0.00)	0.11 (0.01)
Causal effect	-0.05 (0.03)	0.02 (0.02)	0.00 (0.01)	0.01 (0.01)	0.05 (0.02)
$\|nl\|$	5	6	7	8	9
a=1	0.10 (0.01)	0.06 (0.01)	0.08 (0.01)	0.05 (0.01)	0.12 (0.02)
a=0	0.07 (0.01)	0.06 (0.01)	0.06 (0.01)	0.06 (0.00)	0.18 (0.01)
Causal effect	0.03 (0.01)	-0.00 (0.01)	0.02 (0.01)	-0.01 (0.01)	-0.06 (0.02)

Table B.4: The estimated probability of having unanimity (when $|nl| = 0, 9$) or having dissension when the case is about civil rights ($a = 1$) or others ($a = 0$) according to the number of liberal-side vote $|nl|$. The probability of unanimity towards liberal opinions decreases and 5(conservative)-to-4(liberal) decisions increase in this case.

Economic activity

$\|nl\|$	0	1	2	3	4
a=1	0.23 (0.03)	0.09 (0.01)	0.07 (0.01)	0.05 (0.01)	0.07 (0.01)
a=0	0.21 (0.01)	0.10 (0.01)	0.07 (0.00)	0.07 (0.00)	0.13 (0.01)
Causal effect	0.02 (0.03)	-0.01 (0.01)	0.00 (0.01)	-0.01 (0.01)	-0.06 (0.01)
$\|nl\|$	5	6	7	8	9
a=1	0.05 (0.01)	0.05 (0.01)	0.06 (0.01)	0.08 (0.01)	0.25 (0.03)
a=0	0.09 (0.01)	0.06 (0.01)	0.07 (0.01)	0.05 (0.00)	0.15 (0.01)
Causal effect	-0.03 (0.01)	-0.02 (0.01)	-0.00 (0.01)	0.03 (0.01)	0.10 (0.03)

Table B.5: The estimated probability of having unanimity (when $|nl| = 0, 9$) or having dissension when the case is about economic activity($a = 1$) or others ($a = 0$) according to the number of liberal-side vote $|nl|$. The probability of unanimity towards liberal opinions increases in this case.

Judicial power

$\|nl\|$	0	1	2	3	4
a=1	0.32 (0.03)	0.11 (0.01)	0.07 (0.01)	0.05 (0.01)	0.07 (0.01)
a=0	0.19 (0.01)	0.10 (0.01)	0.07 (0.00)	0.07 (0.00)	0.13 (0.01)
Causal effect	0.13 (0.04)	0.01 (0.01)	-0.00 (0.01)	-0.02 (0.01)	-0.06 (0.01)
$\|nl\|$	5	6	7	8	9
a=1	0.06 (0.01)	0.05 (0.01)	0.06 (0.01)	0.06 (0.01)	0.16 (0.03)
a=0	0.08 (0.01)	0.06 (0.01)	0.07 (0.01)	0.06 (0.00)	0.17 (0.01)
Causal ffect	-0.03 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.00 (0.01)	-0.01 (0.03)

Table B.6: The estimated probability of having unanimity (when $|nl| = 0, 9$) or having dissension when the case is about judicial power ($a = 1$) or others ($a = 0$) according to the number of liberal-side vote $|nl|$. The probability of unanimity towards conservative opinions increases and relatively, 5(conservative)-to-4(liberal) decisions decreases.

B.3.2 Simulation using Supreme Court example

To illustrate how chain graphs can be used to estimate causal effects with individual-level treatments, we simulated data from the undirected component of the graph in Figure B.5 with the addition of individual-level treatments A

and individual level covariates C that are dependent across justices and have direct casual effects on A and Y. Treatment A_i nudges Justice i towards a liberal decision. The graph for this setting is too complicated to be helpful, but Figure B.7 illustrates a simplified chain graph following our data generating process for three justices sharing two network ties (between 1 and 2 and between 2 and 3). We specified baseline main effects and pairwise interaction terms using estimates from a log-linear model fit to the actual Supreme Court data, and then varied the magnitude of the main effects and two-way interaction terms by controlling the parameters α and β respectively. For each combination of parameter values, we generated 500 simulated data sets from the chain graph model, each of which used Gibbs sampling to produce 2000 observations of $(\mathbf{Y}, \mathbf{A}, \mathbf{C})$.

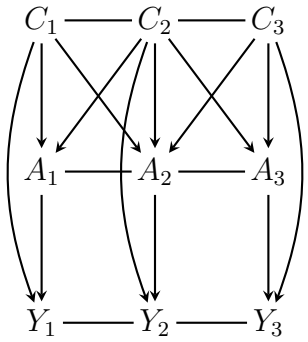


Figure B.7: Simplified 3-node chain graph representing the data generating process used in simulations.

Using the same coefficients of $\hat{h} = \{\hat{h}_i;\ i = 1, 2, \ldots, 9\}$ and $\hat{k} = \{\hat{k}_{ij};\ i, j = 1, 2, \ldots, 9, e_{ij} = 1\}$ from Equation 5.18, a chain component of $(\mathbf{C}, \mathbf{A}, \mathbf{Y})$ are sim-

ulated using Gibbs sampler (Algorithm 2).

Algorithm 2 Simulation data of $(\mathbf{C}, \mathbf{A}, \mathbf{Y})$ using Gibbs sampler.

Data: For s = 0, generate initial values $(\mathbf{C}^{(0)}, \mathbf{A}^{(0)}, \mathbf{Y}^{(0)})$:
begin
$C_i^{(0)} \overset{i.i.d.}{\sim} \mathcal{B}(0.5)\ i = 1, 2, \ldots, 9$;
$A_i^{(0)} \overset{i.i.d.}{\sim} \mathcal{B}(0.3)\ i = 1, 2, \ldots, 9$;
$Y_i^{(0)} \sim \mathcal{B}(\text{logistic}(\hat{h}_i)),\ i = 1, 2, \ldots, 9$;
$Y_i^{(0)} \leftarrow 2Y_i^{(0)} - 1,\ i = 1, 2, \ldots, 9$;
for *s = 0,1, . . . , 2999* **do**
$i \leftarrow \texttt{sample}(\{1, 2, \ldots, 9\})$;
$C_i^{(s+1)} \sim \mathcal{B}\left(\text{logistic}g(C_i|\mathbf{C}_{-i}^{(s)})\right)$;
$A_i^{(s+1)} \sim \mathcal{B}\left(\text{logistic}g(A_i|\mathbf{C}^{(s)}, \mathbf{A}_{-i}^{(s)})\right)$;
$Y_i^{(s+1)} \sim \mathcal{B}\left(\text{logistic}g(Y_i|\mathbf{C}^{(s)}, \mathbf{A}^{(s)}, \mathbf{Y}_{-i}^{(s)})\right)$;
$Y_i^{(s+1)} \leftarrow 2Y_i^{(s+1)} - 1$;
Set $\mathbf{C}_{-i}^{(s+1)} = \mathbf{C}_{-i}^{(s)}$, $\mathbf{A}_{-i}^{(s+1)} = \mathbf{A}_{-i}^{(s)}$, and $\mathbf{Y}_{-i}^{(s+1)} = \mathbf{Y}_{-i}^{(s)}$.
Result: $\{(\mathbf{C}^{(s)}, \mathbf{A}^{(s)}, \mathbf{Y}^{(s)});\ s = 1001, 1002, \ldots, 3000\}$

The above Gibbs sampling utilizes the last $n = 2000$ sequences of $(\mathbf{C}, \mathbf{A}, \mathbf{Y})$ excluding first 1000 burn-in. The Equations B.9 are the conditional densities used in the Gibbs sampling.

$$
\begin{aligned}
g(C_i|\mathbf{C}_{-i}) &= -0.5 - 0.2 \sum_{j \in N(i)\setminus\{i\}} C_j \\
g(A_i|\mathbf{C}, \mathbf{A}_{-i}) &= -0.5 - 0.2 \sum_{j \in N(i)\setminus\{i\}} C_j + 0.1 \sum_{j \in N(i)\setminus\{i\}} A_j \\
g(Y_i|\mathbf{C}, \mathbf{A}, \mathbf{Y}_{-i}) &= \alpha\hat{h}_i + 0.5A_i - 0.2C_i + \beta \sum_{j \in N(i)\setminus\{i\}} \hat{k}_{ij} Y_j.
\end{aligned}
\tag{B.9}
$$

Here the coefficients of $\{\hat{h}_i;\ i = 1, 2, \ldots, 9\}$ and $\{\hat{k}_{ij};\ i, j = 1, 2, \ldots, 9,\ e_{ij} = 1\}$ are from the fitted parameters of the log-linear model (Equation B.10) using

893 decisions made in Supreme Court. It is believed that incorporating such parameters in the simulated data reflects the (relative) magnitude of the main effect and two-way interaction effect embedded in the real data.

$$p\big(\mathbf{Y} = (y_1, y_2, \ldots, y_9)\big) = \frac{1}{Z} \exp\left\{ \sum_{i=1}^{9} h_i y_i + \sum_{i,j=1, e_{ij}=1}^{9} k_{ij} y_i y_j \right\} \tag{B.10}$$

For each of $nr = 500$ simulations having $n = 2000$ observations generated from a Gibbs sampler chain graph model, we fit the (correctly-specified and most parsimonious) log-linear models to estimate unknown parameters in the conditional density of $f(\mathbf{Y}|\mathbf{a}, \mathbf{c};\ \Theta)$. Then by Besag (1974), conditional densities (or conditional clique potentials) of $f(Y_i|\mathbf{Y}_{-i}, \mathbf{a}, \mathbf{c}; \hat{\Theta})$ can be derived for $i = 1, 2, ..., 9$. Details on how to derive conditional densities from the log-linear models using a chain graph are described in Tchetgen et al. (2017). These conditional densities are used for Algorithm 3 to generate counterfactual outcomes.

Let $\mathbf{C}^{(s+1000)} = \mathbf{c}_{(m)}$ for $m = 1, 2, \ldots, n = 2000$ from 2000 sequences of previous Algorithm 3. Probability associated with the counterfactual collective outcome under the treatment a can be estimated using the (hypothetical) ob-

Algorithm 3 Generating Gibbs sampler based on the estimated coefficients.

Data: Intervention vector $\mathbf{a} = (a_1, a_2, \ldots, a_9)$, a set of observed covariates $\mathbf{c} = \{(c_1, c_2, \ldots, c_9)\}$, and estimated coefficients $\hat{\Theta}$.

begin

For s = 0, generate initial values $\mathbf{Y}^{(0)}$:

$Y_i^{(0)} \sim \mathcal{B}(0.5),\ i = 1, 2, \ldots, 9$

$Y_i^{(0)} \leftarrow 2Y_i^{(0)} - 1,\ i = 1, 2, \ldots, 9$

for *s = 0,1, . . . , 5999* **do**

$i \leftarrow \texttt{sample}(\{1, 2, \ldots, 9\})$;

$Y_i^{(s+1)} \sim f(Y_i | \mathbf{c}, \mathbf{a}, \mathbf{Y}_{-i}^{(s)}; \hat{\Theta})$;

$Y_i^{(s+1)} \leftarrow 2Y_i^{(s+1)} - 1$;

Set $\mathbf{Y}_{-i}^{(s+1)} = \mathbf{Y}_{-i}^{(s)}$

Result: $\{\mathbf{Y}^{(s)} = (Y_1^{(s)}, Y_2^{(s)}, \ldots, Y_9^{(s)});\ s = 1001, 1002, \ldots, 6000\}$

served covariates $\{\mathbf{c}_{(m)} = (c_{(m)1}, c_{(m)2}, \ldots, c_{(m)9}) : m = 1, 2, \ldots, n\}$:

$$\begin{aligned} \hat{p}[\mathbf{Y}(\mathbf{a}) = \mathbf{y}] &= \sum_{m=1}^{n} \hat{p}\left(\mathbf{Y}(\mathbf{a}) = \mathbf{y} | \mathbf{a}, \mathbf{c}_{(m)}; \hat{\Theta}\right)/n \\ &= \sum_{m=1}^{n} \left\{ \sum_{s=1001}^{6000} \mathbf{I}\big(\mathbf{Y}^{(s)}(\mathbf{a}, \mathbf{c}_{(m)}; \hat{\Theta}) = \mathbf{y}\big)/5000 \right\}/n. \end{aligned} \tag{B.11}$$

Intervened justice (a)	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 9)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 0)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 5)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 4)$
O'Connor, Scalia, Kennedy, Thomas	0.3672	0.1149	0.0628	0.0646
Stevens, Souter, Ginsburg, Breyer	0.2196	0.0685	0.1216	0.1882
Rehnquist	0.1690	0.2338	0.0679	0.1040
Thomas	0.1717	0.2374	0.0699	0.1066
Stevens	0.1415	0.1957	0.0851	0.1344
Scalia	0.1696	0.2345	0.0703	0.1061

Table B.7: Probability of having unanimous liberal decision ($\sum \mathbf{y} = 9$), unanimous conservative decision ($\sum \mathbf{y} = 0$), five-liberal votes ($\sum \mathbf{y} = 5$), and five-conservative votes ($\sum \mathbf{y} = 4$) under six different treatment assignments. The first set of justices (O'Connor, Scalia, Kennedy, and Thomas) represents conservative arms; while the second set (Stevens, Souter, Ginsburg, and Breyer) represents liberal arms. Rehnquist is a cheif Supreme Court justice; Thomas is known as the most conservative justice among nine while Stevens is the most liberal; Scalia is relatively neutral.

Table B.7 presents the true probability of four different counterfactual out-

comes when $\alpha = \beta = 2$, under six different treatment assignments (treating four conservative justices; treating four liberal justices; treating chief justice Rehnquist; treating Justice Thomas; treating Justice Stevens; treating Justice Scalia). The table shows that in our simulated data treating the four conservative justices results in a higher probability of unanimous liberal decisions (0.37) than treating the four liberal justices (0.22). Similarly, treating the most liberal justice (Stevens) has the smallest effect on the probability of unanimous liberal decisions compared to treating Justices Rehnquist, Thomas, or Scalia. Treating Justice Stevens has the greatest impact on the probabilities of 5-to-4 or 4-to-5 decisions. Treating the four conservative justices together results in relatively high probability of 5(liberal)-to-4(conservative) decisions (0.12).

Figure B.8 and B.9 compare the true probability of counterfactual unanimous votes and neck-and-neck votes respectively and their estimates based on the Gibbs samplers under two treatment assignments (treating four conservative justices and treating four liberal justices). Bias in each estimate and its coverage rate of 95% empirical confidence intervals are presented from Table B.8 to Table B.11.

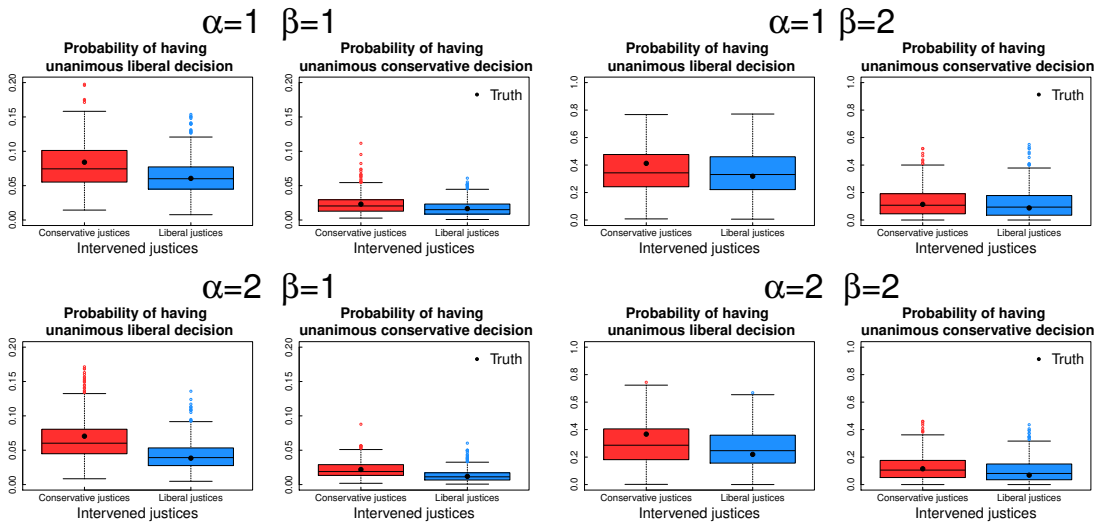


Figure B.8: As α or β increases, we are more likely to have concentrated observations in the certain cell and have less (or empty) observations in the others, so we have a bias and less coverage rates with finite sample $(n = 2000)$. Note that under $\beta = 2$, compared to $\beta = 1$, we observe higher probability of having unanimous decisions. In overall, treating conservative arms is more beneficial than treating liberal arms to draw unanimous liberal decision.

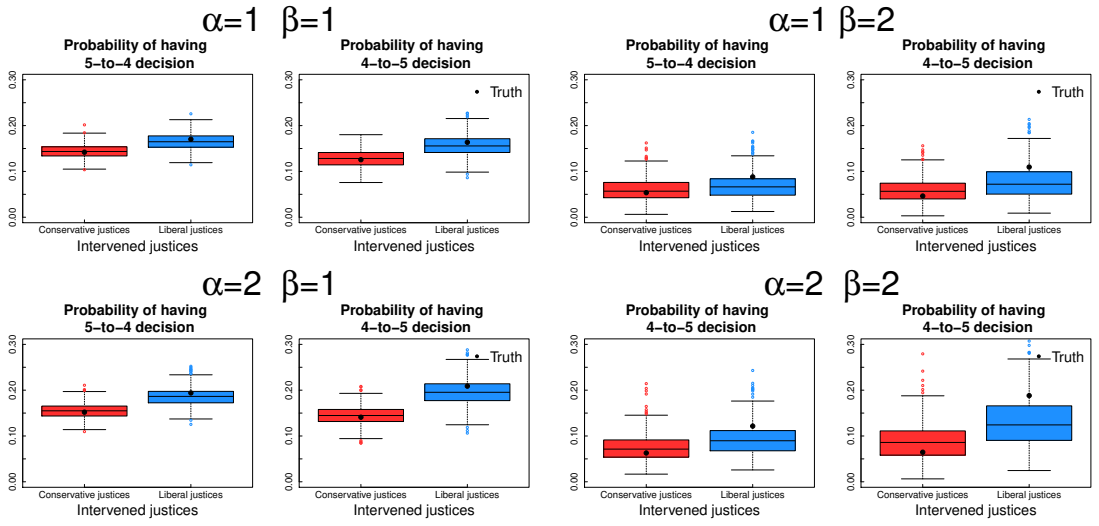


Figure B.9: Generally speaking, treating liberal arms (toward liberal opinion) increases the probability of 5-to-4 or 4-to-5 decisions.

Intervened justice (a)	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 9)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 0)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 5)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 4)$
O'Connor, Scalia, Kennedy, Thomas	-0.0045 (97.00%)	0.0001 (96.00%)	0.0019 (94.20%)	0.0024 (95.40%)
Stevens, Souter, Ginsburg, Breyer	0.0022 (94.40%)	0.0008 (94.40%)	-0.0050 (95.00%)	-0.0069 (93.60%)
Rehnquist	0.0041 (94.60%)	-0.0079 (96.60%)	0.0049 (92.80%)	0.0011 (96.00%)
Thomas	0.0049 (94.00%)	-0.0086 (95.40%)	0.0053 (94.00%)	-0.0000 (95.40%)
Stevens	0.0072 (93.40%)	-0.0081 (95.60%)	0.0050 (92.60%)	-0.0021 (95.40%)
Scalia	0.0062 (94.20%)	-0.0087 (96.00%)	0.0052 (93.80%)	-0.0005 (94.60%)

Table B.8: When $\alpha = 1$ and $\beta = 1$ in Equation B.9, bias and coverage rate of 95% confidence intervals in $nr = 500$ estimated $\hat{p}(\mathbf{Y}(\mathbf{a}) = \mathbf{y})$ assuming correctly specified chain graph.

Intervened justice (a)	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 9)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 0)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 5)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 4)$
O'Connor, Scalia, Kennedy, Thomas	-0.0531 (95.20%)	0.0167 (94.00%)	0.0063 (96.00%)	0.0119 (92.80%)
Stevens, Souter, Ginsburg, Breyer	0.0216 (96.00%)	0.0313 (92.80%)	-0.0191 (91.80%)	-0.0327 (84.40%)
Rehnquist	0.0857 (88.80%)	-0.0963 (85.20%)	0.0081 (93.20%)	0.0024 (94.80%)
Thomas	0.0878 (89.20%)	-0.0970 (85.80%)	0.0077 (93.80%)	0.0017 (94.80%)
Stevens	0.1111 (86.20%)	-0.0890 (86.60%)	0.0019 (95.20%)	-0.0082 (96.00%)
Scalia	0.0898 (89.20%)	-0.0965 (86.20%)	0.0076 (94.60%)	0.0013 (95.00%)

Table B.9: When $\alpha = 1$ and $\beta = 2$ in Equation B.9, bias and coverage rate of 95% confidence intervals in $nr = 500$ estimated $\hat{p}(\mathbf{Y}(\mathbf{a}) = \mathbf{y})$ assuming correctly specified chain graph. Coverage rate when we only intervene a single justice (Rehnquist, Thomas, Stevens, and Scalia) drops due to small number of uanimous observations under a single treatment.

Intervened justice (a)	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 9)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 0)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 5)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 4)$
O'Connor, Scalia, Kennedy, Thomas	-0.0046 (95.00%)	-0.0002 (95.80%)	0.0026 (95.20%)	0.0039 (93.40%)
Stevens, Souter, Ginsburg, Breyer	0.0043 (94.80%)	0.0012 (94.80%)	-0.0079 (92.20%)	-0.0134 (91.60%)
Rehnquist	0.0033 (93.80%)	-0.0053 (96.80%)	0.0044 (94.60%)	0.0008 (95.00%)
Thomas	0.0045 (92.40%)	-0.0060 (96.40%)	0.0041 (94.80%)	-0.0015 (94.40%)
Stevens	0.0059 (92.40%)	-0.0041 (97.00%)	0.0035 (94.00%)	-0.0051 (94.40%)
Scalia	0.0045 (94.20%)	-0.0057 (96.80%)	0.0039 (95.20%)	-0.0016 (95.00%)

Table B.10: When $\alpha = 2$ and $\beta = 1$ in Equation B.9, bias and coverage rate of 95% confidence intervals in $nr = 500$ estimated $\hat{p}(\mathbf{Y}(\mathbf{a}) = \mathbf{y})$ assuming correctly specified chain graph.

Intervened justice (a)	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 9)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 0)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 5)$	$P(\mathbf{Y}(\mathbf{a}) = \mathbf{y}; \sum \mathbf{y} = 4)$
O'Connor, Scalia, Kennedy, Thomas	-0.0682 (93.60%)	0.0087 (94.40%)	0.0125 (92.40%)	0.0236 (90.60%)
Stevens, Souter, Ginsburg, Breyer	0.0402 (94.20%)	0.0333 (91.80%)	-0.0285 (87.60%)	-0.0574 (80.80%)
Rehnquist	0.0646 (91.40%)	-0.0787 (88.80%)	0.0101 (91.60%)	0.0043 (94.60%)
Thomas	0.0676 (91.40%)	-0.0822 (87.60%)	0.0084 (93.80%)	0.0018 (95.80%)
Stevens	0.0935 (88.20%)	-0.0605 (93.60%)	-0.0008 (96.40%)	-0.0163 (96.00%)
Scalia	0.0671 (91.60%)	-0.0784 (90.20%)	0.0088 (93.40%)	0.0025 (95.40%)

Table B.11: When $\alpha = 2$ and $\beta = 2$ in Equation B.9, bias and coverage rate of 95% confidence intervals in $nr = 500$ estimated $\hat{p}(\mathbf{Y}(\mathbf{a}) = \mathbf{y})$ assuming correctly specified chain graph. Similar to Table B.9, the magnitude of two-way interactions (β) that is as strong as real data, may engender almost zero unanimous observations when only a single justice were treated.

Bibliography

Adamic, L., Buyukkokten, O., and Adar, E. (2003). A social network caught in the web. *First monday*, 8(6).

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

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Alexander-Bloch, A. F., Vertes, P. E., Stidd, R., Lalonde, F., Clasen, L., Rapoport, J., Giedd, J., Bullmore, E. T., and Gogtay, N. (2012). The anatomical distance of functional connections predicts brain network topology in health and schizophrenia. *Cerebral cortex*, 23(1):127–138.

Altfeld, M. F. and Spaeth, H. J. (1984). Measuring influence on the us supreme court. *Jurimetrics*, 24(3):236–247.

BIBLIOGRAPHY

Anselin, L., Bera, A. K., Florax, R., and Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional science and urban economics*, 26(1):77–104.

Aral, S., Muchnik, L., and Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549.

Aral, S. and Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

Aral, S. and Walker, D. (2014). Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science*, 60(6):1352–1370.

Aronow, P. M. and Samii, C. (2012). Estimating average causal effects under general interference. Technical report.

Aronow, P. M. and Samii, C. (2013). Estimating average causal effects under interference between units. *arXiv preprint arXiv:1305.6156*.

Athey, S., Eckles, D., and Imbens, G. W. (2016). Exact p-values for network interference*. *Journal of the American Statistical Association*, (just-accepted).

Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network

interference. *Journal of the American Statistical Association*, 113(521):230–240.

Au, R., Massaro, J. M., Wolf, P. A., Young, M. E., Beiser, A., Seshadri, S., D'Agostino, R. B., and DeCarli, C. (2006). Association of white matter hyperintensity volume with decreased cognitive functioning: the framingham heart study. *Archives of neurology*, 63(2):246–250.

Bahr, D. B. and Passerini, E. (1998). Statistical mechanics of opinion formation and collective behavior: Micro-sociology. *The Journal of mathematical sociology*, 23(1):1–27.

Bailey, N. T. et al. (1975). *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417.

BIBLIOGRAPHY

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2013). The diffusion of microfinance. *Science*, 341(6144):1236498.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. (2014). Gossip: Identifying central individuals in a social network. Technical report, National Bureau of Economic Research.

Bauch, C. T. and Galvani, A. P. (2013). Social factors in epidemiology. *Science*, 342(6154):47–49.

Beaman, L., BenYishay, A., Magruder, J., and Mobarak, A. M. (2015). Can network theory based targeting increase technology adoption. *Unpublished manuscript*.

Berkman, L. and Syme, S. (1979). Social networks, host resistance, and mortality: a nine-year follow-up study of alameda county residents. *American journal of Epidemiology*, 109(2):186.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The statistician*, pages 179–195.

Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., and

Walczak, A. M. (2012). Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*.

Binney, J. J., Dowrick, N. J., Fisher, A. J., and Newman, M. (1992). *The theory of critical phenomena: an introduction to the renormalization group*. Oxford University Press, Inc.

Black, W. R. (1992). Network autocorrelation in transport network and flow systems. *Geographical Analysis*, 24(3):207–222.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298.

Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, 27(1):55–71.

Bowers, J., M, F. M., and C, P. (2013). Reasoning about interference between units: A general framework. *Political Analysis*, 21:97–124.

Butts, C. T. et al. (2008). Social network analysis with sna. *Journal of Statistical Software*, 24(6):1–51.

BIBLIOGRAPHY

Cai, J., De Janvry, A., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

Callen, H. B. (1998). Thermodynamics and an introduction to thermostatistics.

Castellano, C. (2012). Social influence and the dynamics of opinions: the approach of statistical physics. *Managerial and Decision Economics*, 33(5-6):311–321.

Castellano, C., Fortunato, S., and Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591.

Castelli, W. (1988). Cholesterol and lipids in the risk of coronary artery disease–the framingham heart study. *The Canadian journal of cardiology*, 4:5A–10A.

Centola, D. (2010). The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197.

Centola, D. (2011). An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272.

Chami, G. F., Ahnert, S. E., Voors, M. J., and Kontoleon, A. A. (2014). Social network analysis predicts health behaviours and self-reported health in african villages. *PloS one*, 9(7):e103500.

BIBLIOGRAPHY

Chandler, D. (1987). Introduction to modern statistical mechanics. *Introduction to Modern Statistical Mechanics, by David Chandler, pp. 288. Foreword by David Chandler. Oxford University Press, Sep 1987. ISBN-10: 0195042778. ISBN-13: 9780195042771*, page 288.

Chaudhuri, S. and Richardson, T. (2002). Using the structure of d-connecting paths as a qualitative measure of the strength of dependence. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 116–123. Morgan Kaufmann Publishers Inc.

Chen, B. L., Hall, D. H., and Chklovskii, D. B. (2006). Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, 103(12):4723–4728.

Chen, L., Shen, C., Vogelstein, J. T., and Priebe, C. E. (2016). Robust vertex classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):578–590.

Chen, W., Wang, C., and Wang, Y. (2010). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM.

Chen, W., Wang, Y., and Yang, S. (2009). Efficient influence maximization in

social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM.

Cherniak, C., Mokhtarzada, Z., Rodriguez-Esteban, R., and Changizi, K. (2004). Global optimization of cerebral cortex layout. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4):1081–1086.

Chin, A., Eckles, D., and Ugander, J. (2018). Evaluating stochastic seeding strategies in networks. *arXiv preprint arXiv:1809.09561*.

Choi, D. S. (2014). Estimation of monotone treatment effects in network experiments. *arXiv preprint arXiv:1408.4102*.

Christakis, N. and Fowler, J. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379.

Christakis, N. and Fowler, J. (2008). The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258.

Christley, R. M., Pinchbeck, G., Bowers, R., Clancy, D., French, N., Bennett, R., and Turner, J. (2005). Infection in social networks: using network analysis to identify high-risk individuals. *American journal of epidemiology*, 162(10):1024–1031.

BIBLIOGRAPHY

Cliff, A. and Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical analysis*, 4(3):267–284.

Cliff, A. D. and Ord, J. K. (1968). *The problem of spatial autocorrelation*. University of Bristol, Department of Economics and Department of Geography.

Cliff, A. D. and Ord, K. (1970). Spatial autocorrelation: a review of existing and new measures with applications. *Economic Geography*, 46(sup1):269–292.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431.

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer.

da Silva, E. C., Silva, A. C., de Paiva, A. C., and Nunes, R. A. (2008). Diagnosis of lung nodule using moran's index and geary's coefficient in computerized tomography images. *Pattern Analysis and Applications*, 11(1):89–99.

D'Agostino, R. B., Russell, M. W., Huse, D. M., Ellison, R. C., Silbershatz, H., Wilson, P. W., and Hartz, S. C. (2000). Primary and subsequent coronary risk

appraisal: new results from the framingham study. *American heart journal*, 139(2):272–281.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., and Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care the framingham heart study. *Circulation*, 117(6):743–753.

Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *The Annals of Probability*, pages 745–764.

Diniz-Filho, J. A. F., Bini, L. M., and Hawkins, B. A. (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global ecology and Biogeography*, 12(1):53–64.

Eckles, D., Karrer, B., and Ugander, J. (2014). Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530*.

Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180.

F Dormann, C., M McPherson, J., B Araújo, M., Bivand, R., Bolliger, J., Carl, G., G Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., et al. (2007). Methods

to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628.

Farber, S., Marin, M. R., and Páez, A. (2015). Testing for spatial independence using similarity relations. *Geographical Analysis*, 47(2):97–120.

Farber, S., Páez, A., and Volz, E. (2009). Topology and dependency tests in spatial and network autoregressive models. *Geographical Analysis*, 41(2):158–180.

Forastiere, L., Airoldi, E. M., and Mealli, F. (2016). Identification and estimation of treatment and interference effects in observational studies on networks. *arXiv preprint arXiv:1609.06245*.

Fortin, M.-J., Drapeau, P., and Legendre, P. (1989). Spatial autocorrelation and sampling design in plant ecology. *Vegetatio*, 83(1-2):209–222.

Fosdick, B. K. and Hoff, P. D. (2015). Testing and modeling dependencies between a network and nodal attributes. *Journal of the American Statistical Association*, 110(511):1047–1056.

Fouss, F., Saerens, M., and Shimbo, M. (2016). *Algorithms and Models for Network Data and Link Analysis*. Cambridge University Press.

Fowler, J. H. and Christakis, N. A. (2008). Dynamic spread of happiness in a

large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj*, 337:a2338.

Fowler, J. H. and Christakis, N. A. (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, 107(12):5334–5338.

Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S., and Wahlbeck, P. J. (2007). Network analysis and the law: Measuring the legal importance of precedents at the us supreme court. *Political Analysis*, 15(3):324–346.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146.

Getis, A. and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3):189–206.

Gibbs, J. W. (2014). *Elementary principles in statistical mechanics*. Courier Corporation.

Goldenberg, J., Libai, B., and Muller, E. (2001). Using complex systems analysis to advance marketing theory development: Modeling heterogeneity ef-

fects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review*, 2001:1.

Gordon, T., Castelli, W. P., Hjortland, M. C., Kannel, W. B., and Dawber, T. R. (1977). High density lipoprotein as a protective factor against coronary heart disease: the framingham study. *The American journal of medicine*, 62(5):707–714.

Grabowski, A. and Kosiński, R. (2006). Ising-based model of opinion formation in a complex network of interpersonal interactions. *Physica A: Statistical Mechanics and its Applications*, 361(2):651–664.

Graham, B., Imbens, G., and Ridder, G. (2010). Measuring the effects of segregation in the presence of social spillovers: A nonparametric approach. Technical report, National Bureau of Economic Research.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443.

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, pages 300–306.

Gretton, A. and Gyorfi, L. (2010). Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423.

BIBLIOGRAPHY

Hanneke, S. and Xing, E. P. (2009). Network completion and survey sampling. In *Artificial Intelligence and Statistics*, pages 209–215.

Heller, R., Heller, Y., and Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510.

Heller, R., Heller, Y., Kaufman, S., Brill, B., and Gorfine, M. (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(29):1–54.

Hernandez-Hernandez, G., Myers, J., Alvarez-Lacalle, E., and Shiferaw, Y. (2017). Nonlinear signaling on biological networks: The role of stochasticity and spectral clustering. *Physical Review E*, 95(3):032313.

Hong, G. and Raudenbush, S. (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, 101(475):901–910.

Hong, G. and Raudenbush, S. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, 33(3):333–362.

Howard, M., Cox Pahnke, E., Boeker, W., et al. (2016). Understanding network formation in strategy research: Exponential random graph models. *Strategic Management Journal*, 37(1):22–44.

BIBLIOGRAPHY

Huckfeldt, R. R. and Sprague, J. (1995). *Citizens, politics and social communication: Information and influence in an election campaign*. Cambridge University Press.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Ilyas, M. U. and Radha, H. (2011). Identifying influential nodes in online social networks using principal component centrality. In *Communications (ICC), 2011 IEEE International Conference on*, pages 1–5. IEEE.

Jagadeesan, R., Pillai, N., and Volfovsky, A. (2017). Designs for estimating the treatment effect in networks with interference. *arXiv preprint arXiv:1705.08524*.

Kaiser, M. and Hilgetag, C. C. (2006). Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. *PLoS computational biology*, 2(7):e95.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107.

Katona, Z., Zubcsek, P. P., and Sarvary, M. (2011). Network effects and personal

influences: The diffusion of an online social network. *Journal of marketing research*, 48(3):425–443.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kaufman, J. S. (2017). *Methods in social epidemiology*, volume 16. John Wiley & Sons.

Kawachi, I. and Berkman, L. F. (2001). Social ties and mental health. *Journal of Urban health*, 78(3):458–467.

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM.

Kenny, C. (1998). The behavioral consequences of political discussion: Another look at discussant effects on vote choice. *the Journal of Politics*, 60(1):231–244.

Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O'Malley, A. J., Fowler, J. H., and Christakis, N. A. (2015). Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 386(9989):145–153.

BIBLIOGRAPHY

Kiss, C. and Bichler, M. (2008). Identification of influencersmeasuring influence in customer networks. *Decision Support Systems*, 46(1):233–253.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893.

Klemm, K., Serrano, M. Á., Eguíluz, V. M., and San Miguel, M. (2012). A measure of individual role in collective dynamics. *Scientific reports*, 2:292.

Kosma, M. N. (1998). Measuring the influence of supreme court justices. *The Journal of Legal Studies*, 27(2):333–372.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *science*, 311(5757):88–90.

Lafon, S. and Lee, A. B. (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE transactions on pattern analysis and machine intelligence*, 28(9):1393–1403.

Lam, N. S.-N., Qiu, H.-l., Quattrochi, D. A., and Emerson, C. W. (2002). An evaluation of fractal methods for characterizing image complexity. *Cartography and Geographic Information Science*, 29(1):25–35.

Lauer, M. S., Anderson, K. M., Kannel, W. B., and Levy, D. (1991). The impact of

obesity on left ventricular mass and geometry: the framingham heart study. *Jama*, 266(2):231–236.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, U.K.: Clarendon.

Lauritzen, S. L. and Richardson, T. S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348.

Lee, E. D., Broedersz, C. P., and Bialek, W. (2015). Statistical mechanics of the us supreme court. *Journal of Statistical Physics*, 160(2):275–301.

Lee, Y. and Ogburn, E. L. (2018a). *netdep: Testing for Network Dependence*. R package version 0.1.0.

Lee, Y. and Ogburn, E. L. (2018b). Testing for network and spatial autocorrelation. *arXiv preprint arXiv:1710.03296*.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74(6):1659–1673.

Lennon, J. J. (2000). Red-shifts and red herrings in geographical ecology. *Ecography*, 23(1):101–113.

Levy, D., Garrison, R. J., Savage, D. D., Kannel, W. B., and Castelli, W. P. (1990). Prognostic implications of echocardiographically determined left ventricular

mass in the framingham heart study. *New England Journal of Medicine*, 322(22):1561–1566.

Lewis, K., Gonzalez, M., and Kaufman, J. (2012). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1):68–72.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using facebook. com. *Social networks*, 30(4):330–342.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, pages 13–22.

Liang, X., Zou, Q., He, Y., and Yang, Y. (2013). Coupling of functional connectivity and regional cerebral blood flow reveals a physiological basis for network hubs of the human brain. *Proceedings of the National Academy of Sciences*, 110(5):1929–1934.

Lichstein, J. W., Simons, T. R., Shriner, S. A., and Franzreb, K. E. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological monographs*, 72(3):445–463.

Lin, D. Y., Wei, L.-J., and Ying, Z. (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572.

BIBLIOGRAPHY

Liu, L., Hudgens, M., and Becker-Dreps, S. (2016). On inverse probability-weighted estimators in the presence of interference. *Biometrika*, 103(4):829–842.

Liu, L. and Hudgens, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *Journal of the american statistical association*, 109(505):288–301.

Liu, S., Ying, L., and Shakkottai, S. (2010). Influence maximization in social networks: An ising-model-based approach. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 570–576. IEEE.

Long, J., Harre, N., and Atkinson, Q. D. (2015). Social clustering in high school transport choices. *Journal of environmental psychology*, 41:155–165.

Lü, L., Zhou, T., Zhang, Q.-M., and Stanley, H. E. (2016). The h-index of a network node and its relation to degree and coreness. *Nature communications*, 7:10168.

Lu, Y. and Wang, X. (2011). Understanding complex legislative and judicial behaviour via hierarchical ideal point estimation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(1):93–107.

BIBLIOGRAPHY

Lucas, A. (2013). Binary decision making with very heterogeneous influence. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(09):P09024.

Lynn, C. and Lee, D. D. (2016). Maximizing influence in an ising network: A mean-field optimal solution. In *Advances in Neural Information Processing Systems*, pages 2495–2503.

Lyons, R. (2011). The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy*, 2(1).

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220.

Moed, H. F. (2006). *Citation analysis in research evaluation*, volume 9. Springer Science & Business Media.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251.

Narayanam, R. and Narahari, Y. (2011). A shapley value-based approach to discover influential nodes in social networks. *IEEE Transactions on Automation Science and Engineering*, 8(1):130–147.

BIBLIOGRAPHY

Nekovee, M., Moreno, Y., Bianconi, G., and Marsili, M. (2007). Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 374(1):457–470.

Newman, M. (2010). *Networks: an introduction*. Oxford university press.

Newman, M. (2018). *Networks*. Oxford university press.

Nickerson, D. W. (2008). Is voting contagious? evidence from two field experiments. *American Political Science Review*, 102(1):49–57.

Ogburn, E. L. (2017). Challenges to estimating contagion effects from observational data. *arXiv preprint arXiv:1706.08440*.

Ogburn, E. L., Shpitser, I., and Lee, Y. (2018a). Causal inference, social networks, and chain graphs. *arXiv preprint arXiv:1812.04990*.

Ogburn, E. L., Shpitser, I., and Lee, Y. (2018b). Causal inference, social networks, and chain graphs. *arXiv preprint arXiv:1812.04990*.

Ogburn, E. L., Sofrygin, O., Diaz, I., and van der Laan, M. J. (2017). Causal inference for social network data. *arXiv preprint arXiv:1705.08527*.

Ogburn, E. L. and VanderWeele, T. J. (2017). Vaccines, contagion, and social networks. *Annals of Applied Statistcs*.

Ogburn, E. L., VanderWeele, T. J., et al. (2014). Causal diagrams for interference. *Statistical science*, 29(4):559–578.

BIBLIOGRAPHY

O'Neil, K. A. and Redner, R. A. (1993). Asymptotic distributions of weighted u-statistics of degree 2. *The Annals of Probability*, pages 1159–1169.

O'Neill, B. (2009). Exchangeability, correlation and bayes' effect. *International Statistical Review*, 77(2):241250.

Orbanz, P. (2017). Subsampling large graphs and invariance in networks. *arXiv preprint arXiv:1710.04217*.

Orbanz, P. and Roy, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):437–461.

Overmars, K. d., De Koning, G., and Veldkamp, A. (2003). Spatial autocorrelation in multi-scale land use models. *Ecological modelling*, 164(2):257–270.

Pachucki, M. A., Jacques, P. F., and Christakis, N. A. (2011). Social network concordance in food choice among spouses, friends, and siblings. *American Journal of Public Health*, 101(11):2170–2177.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Paluck, E. L., Shepherd, H., and Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113(3):566–571.

BIBLIOGRAPHY

Papadogeorgou, G. (2017). Replication data for: Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching.

Papadogeorgou, G., Choirat, C., and Zigler, C. M. (2016). Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics*.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge Univ Press.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Peel, L., Larremore, D. B., and Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*.

Perez-Heydrich, C., Hudgens, M. G., Halloran, M. E., Clemens, J. D., Ali, M., and Emch, M. E. (2014). Assessing effects of cholera vaccination in the presence of interference. *Biometrics*, 70(3):731–741.

BIBLIOGRAPHY

Perisic, A. and Bauch, C. T. (2009). Social contact networks and disease eradicability under voluntary vaccination. *PLoS computational biology*, 5(2):e1000280.

Pryor, T. (2017). Using citations to measure influence on the supreme court. *American Politics Research*, 45(3):366–402.

Qiu, W. Q., Dean, M., Liu, T., George, L., Gann, M., Cohen, J., and Bruce, M. L. (2010). Physical and mental health of homebound older adults: an overlooked population. *Journal of the American Geriatrics Society*, 58(12):2423–2428.

Rand, D. G., Arbesman, S., and Christakis, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108(48):19193–19198.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *preprint:* `http://www.csss.washington.edu/Papers/wp128.pdf`.

Riggs, R. E. (1992). When every vote counts: 5-4 decisions in the united states supreme court, 1900-90. *Hofstra L. Rev.*, 21:667.

Rizzo, M. and Székely, G. (2016). Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38.

BIBLIOGRAPHY

Robinaugh, D. J., Millner, A. J., and McNally, R. J. (2016). Identifying highly influential nodes in the complicated grief network. *Journal of Abnormal Psychology*, 125(6):747.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915.

Rosenbaum, P. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenquist, J. N., Murabito, J., Fowler, J. H., and Christakis, N. A. (2010). The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, 152(7):426–433.

Rubin, D. (1990a). On the application of probability theory to agricultural experiments. essay on principles. section 9. comment: Neyman (1923) and

causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26.

Rubin, D. B. (1990b). Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Russell, D. W. and Cutrona, C. E. (1991). Social support, stress, and depressive symptoms among the elderly: Test of a process model. *Psychology and aging*, 6(2):190.

Saczynski, J. S., Beiser, A., Seshadri, S., Auerbach, S., Wolf, P., and Au, R. (2010). Depressive symptoms and risk of dementia the framingham heart study. *Neurology*, 75(1):35–41.

Saito, K., Kimura, M., Ohara, K., and Motoda, H. (2012). Efficient discovery of

influential nodes for sis models in social networks. *Knowledge and information systems*, 30(3):613–635.

Saito, K., Kimura, M., Ohara, K., and Motoda, H. (2016). Super mediator–a new centrality measure of node importance for information diffusion over social network. *Information Sciences*, 329:985–1000.

Sen, A. (1976). Large sample-size distribution of statistics used in testing for spatial correlation. *Geographical analysis*, 8(2):175–184.

Shalizi, C. and Thomas, A. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239.

Shapiro, C. P. and Hubert, L. (1979). Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *The Annals of Statistics*, pages 788–794.

Shen, C., Priebe, C. E., and Vogelstein, J. T. (2018a). From distance correlation to multiscale graph correlation. *arXiv preprint arXiv:1710.09768*.

Shen, C., Wang, Q., Bridgeford, E., Priebe, C. E., Maggioni, M., and Vogelstein, J. T. (2018b). Discovering relationships and their structures across disparate data modalities. *https://arxiv.org/abs/1609.05148*.

Sikic, M., Lancic, A., Antulov-Fantulin, N., and Stefancic, H. (2013). Epidemic

centralityis there an underestimated epidemic impact of network peripheral nodes? *The European Physical Journal B*, 86(10):440.

Sillanpää, M. (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity*, 106(4):511.

Simko, G. I. and Csermely, P. (2013). Nodes having a major influence to break cooperation define a novel centrality measure: game centrality. *PloS one*, 8(6):e67159.

Sirovich, L. (2003). A pattern analysis of the second rehnquist us supreme court. *Proceedings of the National Academy of Sciences*, 100(13):7432–7437.

Sison, C. P. and Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, 90(429):366–369.

Smith, S. T., Kao, E. K., Shah, D. C., Simek, O., and Rubin, D. B. (2018). Influence estimation on social media networks using causal inference. *arXiv preprint arXiv:1804.04109*.

Smouse, P. E. and Peakall, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, 82(5):561–573.

BIBLIOGRAPHY

Sobel, M. (2006). What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101(476):1398–1407.

Songer, D. R. and Lindquist, S. A. (1996). Not the whole story: The impact of justices' values on supreme court decision making. *American Journal of Political Science*, 40(4):1049–1063.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Spirtes, P. and Verma, T. (1992). Equivalence of causal models with latent variables.

Sunstein, C. R. (2014). Unanimity and disagreement on the supreme court. *Cornell L. Rev.*, 100:769.

Sussman, D., Tang, M., Fishkind, D., and Priebe, C. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128.

Sussman, D. L., Tang, M., and Priebe, C. E. (2014). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57.

Székely, G. and Rizzo, M. (2013a). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

BIBLIOGRAPHY

Székely, G. and Rizzo, M. (2014). Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, 42(6):2382–2412.

Székely, G. J. and Rizzo, M. L. (2013b). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Tableman, M. and Kim, J. S. (2003). *Survival analysis using S: analysis of time-to-event data*. CRC press.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., and Priebe, C. E. (2017). A nonparametric two-sample hypothesis testing problem for random dot product graphs. *Bernoulli*, 23(3):1599–1630.

Tate, C. N. (1981). Personal attribute models of the voting behavior of us supreme court justices: Liberalism in civil liberties and economics decisions, 1946–1978. *American Political Science Review*, 75(2):355–367.

Tchetgen, E. J. T., Fulcher, I., and Shpitser, I. (2017). Auto-g-computation of causal effects on a network. *arXiv preprint arXiv:1709.01577*.

Tchetgen, E. J. T. and VanderWeele, T. J. (2012). On causal inference in the

presence of interference. *Statistical methods in medical research*, 21(1):55–75.

Tchetgen Tchetgen, E. J. and VanderWeele, T. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75.

Trogdon, J. G., Nonnemaker, J., and Pais, J. (2008). Peer effects in adolescent overweight. *Journal of health economics*, 27(5):1388–1399.

Trusov, M., Bucklin, R. E., and Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of marketing*, 73(5):90–102.

Tsao, C. W. and Vasan, R. S. (2015). Cohort profile: The framingham heart study (fhs): overview of milestones in cardiovascular epidemiology. *International journal of epidemiology*, 44(6):1800–1813.

Tsuji, H., Venditti, F. J., Manders, E. S., Evans, J. C., Larson, M. G., Feldman, C. L., and Levy, D. (1994). Reduced heart rate variability and mortality risk in an elderly cohort. the framingham heart study. *Circulation*, 90(2):878–883.

Valente, T. W. (2012). Network interventions. *Science*, 337(6090):49–53.

BIBLIOGRAPHY

van der Laan, M. J. (2014). Causal inference for a population of causally connected units. *Journal of Causal Inference J. Causal Infer.*, 2(1):13–74.

VanderWeele, T. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, 38(4):515–544.

VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in medicine*, 27(11):1934–1943.

VanderWeele, T. J., Vandenbroucke, J. P., Tchetgen, E. J. T., and Robins, J. M. (2012). A mapping between interactions and interference: implications for vaccine trials. *Epidemiology (Cambridge, Mass.)*, 23(2):285.

Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H., and Chklovskii, D. B. (2011). Structural properties of the caenorhabditis elegans neuronal network. *PLoS computational biology*, 7(2):e1001066.

Vasan, R. S., Pencina, M. J., Cobain, M., Freiberg, M. S., and D'Agostino, R. B. (2005). Estimated risks for developing obesity in the framingham heart study. *Annals of internal medicine*, 143(7):473–480.

Voorhees, C. C., Murray, D., Welk, G., Birnbaum, A., Ribisl, K. M., Johnson, C. C., Pfeiffer, K. A., Saksvig, B., and Jobe, J. B. (2005). The role of peer

social network factors and physical activity in adolescent girls. *American Journal of Health Behavior*, 29(2):183–190.

Wang, P., Lü, J., and Yu, X. (2014). Identification of important nodes in directed biological networks: A network motif approach. *PloS one*, 9(8):e106132.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. *Psychometrika*, 61(3):401–425.

Wolf, P. A., Abbott, R. D., and Kannel, W. B. (1991). Atrial fibrillation as an independent risk factor for stroke: the framingham study. *Stroke*, 22(8):983–988.

Xin, L., Zhu, M., Chipman, H., et al. (2017). A continuous-time stochastic block model for basketball networks. *The Annals of Applied Statistics*, 11(2):553–597.

Zhu, M. and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis*, 51:918–930.

Vita

YOUJIN LEE

ylee160@jhu.edu
615 N.Wolfe Street E3037 ⋄ Baltimore, Maryland 21205
Github: youjin1207 Personal page : http://www.youjinleeylee.com

PROFESSIONAL EXPERIENCE

Post-Doctoral Fellow February 2019 - June 2019 (Expected)
Department of Mental Health, Johns Hopkins School of Public Health
Supervised by Elizabeth A. Stuart

EDUCATION

Doctor of Philosophy in Biostatistics August 2014 - January 2019
Department of Biostatistics, Johns Hopkins School of Public Health
Dissertation title: *Statistical Reasoning in Network Data*
Primary Advisor : Elizabeth L. Ogburn
Thesis Committee : Carl Latkin, Ilya Shpitser, and Abhirup Datta

B.S. with honors in Statistics March 2010 - August 2014
Department of Statistics, Seoul National University, South Korea
Graduated summa cum laude

RELATED EXPERIENCE

Research Assistant August 2015 - January 2019
Johns Hopkins University

Working Group Sep 2014 - Present
Causal Inference Working Group
Survival, Longitudinal, and Multivariate Analysis (SLAM) Working Group

Research Intern June 2012 - August 2012
Bioinformatics and Biostatistics Lab, Seoul National University

RESEARCH INTEREST

Causal inference, Social network, Interference, Respondent-Driven Sampling, Peer effect, Competing-risks.

PAPERS

Ogburn, E. L., Shpitser, I. & **Lee, Y.** (2018). 'Collective problem solving, causal inference, and chain graphs'. arXiv preprint arXiv:1812.04990. *Under Review*

Lee, Y., & Ogburn, E. L. (2018). 'Invalid Statistical Inference Due to Social Network Dependence'. *Under Review*

Lee, Y., Grantz, KL., Wang, MC. & Sundaram, R. (2018), 'Joint Modeling of Competing Risks and Current Status Data: An Application to Spontaneous Labor Study'. *Under Minor Revision for Journal of the Royal Statistical Society, Series C*

Lee, Y., & Ogburn, E. L. (2017). 'Testing for Network and Spatial Autocorrelation'. arXiv preprint arXiv:1710.03296.

Lee, Y., Shen, C., Priebe, CE., & Vogelstein, J.T. (2017) , 'Network Dependence Testing via Diffusion Maps and Distance-Based Correlations', arXiv preprint arXiv:1703.10136. *Under Minor Revision for Biometrika*

Manuscript in Preparation:

- **Lee, Y.**, Shpitser, I., & Ogburn, E. L. (2018+). 'Identifying causally influential subjects on a social network'.
- Liu, L., **Lee, Y.**, Hong, X., Hao, L., Burd, I., Wang, MC. and Wang, X. (2018+), 'A comprehensive array of reproductive history and risk of preterm birth: new insights from the Boston Birth Cohort'
- **Lee, Y.**, Liu, L., & Wang, MC. (2018+), 'Caution in reporting relative risk from logistic regression model'

SOFTWARE

`R` package

- `logisticRR` (author, maintainer) : An `R` package for deriving adjusted relative risks from logistic regression. [CRAN]
- `netdep` (author, maintainer): An `R` package for testing network dependence and generating dependent observations. [CRAN]
- `netchain` (author, maintainer) : An `R` package for estimating probability associated with collective counterfactual outcomes under interference. [CRAN]
- `MGC` (author): An `R` package for investigating relationships between properties of a dataset and the underlying geometries of the relationships.

Computer skills

- `R`, `C++`, `Stata`, `SAS`, LaTeX

PRESENTATION

Talk

- 2019 Conference on Lifetime Data Science: Foundations and Frontiers. May 29-31, 2019; Pittsburgh, PA. (*upcoming*)
- Invalid Statistical Inference Due to Social Network Dependence, Joint Statistical Meetings. July 28 - August 2, 2018; Vancouver, Canada.
- Joint Modeling of Competing Risks and Current Status Data: An Application to Spontaneous Labor Study, Eastern North American Region International Biometric Society. March 25 - March 28, 2018; Atlanta, GA.
- Testing Independence in Network via a family of network metrics, Joint Statistical Meetings. July 29 - August 3, 2017; Baltimore, MD.

Poster

- Collective problem solving, causal inference, and chain graphs, Atlantic Causal Inference Conference 2018, May 22-23, 2018; Carnegie Mellon University
- Joint Modeling of Delivery Time and Onset Time of Morbidities during the Second-stage Labor, 2017 Conference on Lifetime Data Science. May 25-27, 2017; University of Connecticut.

· Testing Independence between Observations from a Single Network, Eastern North American Region International Biometric Society. March 12-15, 2017; Washington, DC.

PROFESSIONAL ACTIVITIES

Reviewer : *Journal of the American Statistical Association*, *Journal of Causal Inference*

Volunteering : Information service at *19th New Researchers Conference (NRC)*

AWARD

The Jane and Steve Dykacz Award 2018
For outstanding paper by a Biostatistics student in the area of medical statistics, Department of Biostatistics, Johns Hopkins School of Public Health

The Margaret Merrell Award 2018
For outstanding research by a Biostatistics doctoral student, Department of Biostatistics, Johns Hopkins School of Public Health

Winner of Student Paper Awards Joint Statistical Meetings (JSM) 2017
ASA Nonparametric Statistics Section

Winner of Student Poster Award Conference on Lifetime Data Science 2017

Louis I. and Thomas D. Dublin Award 2016
For the advancement of Epidemiology and Biostatistics supports for students, Department of Biostatistics, Johns Hopkins School of Public Health

SCHOLARSHIP

Recipient of overseas scholarship, Kwanjeong Educational Foundation 2014-2018

Recipient of National Science and Engineering Scholarship, Korea Student Aid Foundation, Full tuition exemption 2010-2013

TEACHING ASSISTANT

Public Health Biostatistics (Undergraduate Course) Fall 2018
Instructor : Margaret Taub and Leah Jager

Causal Inference in Medicine and Public Health I 2017-2018 3rd and 4th term
Instructor : Elizabeth Stuart
Lecture : Causal inference under interference [slide]

Survival Analysis I-II 2017-2018 1st and 2nd term
Instructor : Mei-Cheng Wang

Survival Analysis Summer 2017
Instructor : Xiangrong Kong
Graduate Summer Institute of Epidemiology and Biostatistics

Causal Inference in Medicine and Public Health I 2016-2017 3rd and 4th term
Instructor : Elizabeth Stuart
Lecture : Introduction to principal stratification and truncation due to death

Statistical Reasoning in Public Health II — 2016-2017 2nd term
Instructor : Marie Diener-West and Karen Bandeen-Roche

Survival Analysis I — 2016-2017 1st term
Instructor : Chiung-Yu Huang

Statistical Reasoning in Public Health IV — 2015-2016 4th term
Instructor : James Tonascia

Statistical Reasoning in Public Health III — 2015-2016 3rd term
Instructor : John McGready and Marie Diener-West

Statistical Reasoning in Public Health I - II — 2015-2016 1st and 2nd term
Instructor : John McGready

OTHER EXPERIENCE

Language Tutoring — March 2013 - June 2013
Faculty of Liberal Education, Seoul National University

Exchange Student Program — August 2012 - December 2012
University of British Columbia, Canada

VITA