# Latent Variable Models for Indirectly or Imprecisely Measured Networks

Wesley Lee

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Tyler H. McCormick, Chair

Adrian Dobra

Elena A. Erosheva

Program Authorized to Offer Degree:
Statistics

University of Washington

**Abstract**

Latent Variable Models for Indirectly or Imprecisely Measured Networks

Wesley Lee

Chair of the Supervisory Committee:
Associate Professor Tyler H. McCormick
Department of Statistics

In many scientific settings, networks are important structures used to represent the relationships between actors in a population of study. The most common methods for measuring networks are to survey study participants about who their connections are and to collect interaction activity between pairs of actors. However, directly measuring the exact network of interest can be challenging. In the context of surveys, participants do not always provide accurate accounts of their connections, which can result in mismeasurement of the network. In context of logged activity data, interactions do not directly quantify relationships between individuals or the propensity to interact in the future. In this thesis, we broadly conceptualize observed data from either source as manifestations from a latent network of interest, and we seek to use the former to infer the structure of the latter.

In Chapter 2, we demonstrate how using mismeasured network data can affect subsequent inference, specifically in the setting of experiments on networks. In these experiments, individuals are not only influenced by their own treatment assignments, but also by those of their peers; these indirect treatment effects are often of direct scientific interest. In order to measure these indirect effects, researchers typically collect network data by surveying subjects about their connections. However, both survey design decisions and misreporting can lead to an observed network with mismeasurement. We show that mismeasured connections can in turn bias existing estimators of treatment effects, but this bias can be attenuated

by explicitly accounting for (via a mixture model) the relationship between the observed, mismeasured network and the latent network of interest.

An alternate source of network data to surveys are relational event data, consisting of interactions between pairs of actors over time. Typically recorded using automated data-gathering technology, relational event data can potentially sidestep design and misreporting issues more common in survey data but present their own additional modelling challenges. These events are typically measured in continuous-time and do not directly quantify relationships between actors, preventing their direct use in inference problems such as the experimental setting considered in Chapter 2. We propose a continuous-time point-process model for inferring a network of social relations from interaction data in Chapter 3. We allow the propensity for interactions to depend on time and covariates, in addition to the dynamic latent network, thus decoupling observed interaction counts from relational strength.

In Chapter 4, we address another issue with modeling relational event data: the potentially large scale of the networks on which the data is collected. As data-gathering technology becomes more ubiquitous, relational event data is able to measure activity on networks of a much larger size than accessible via survey sources. Estimation for many existing models becomes computationally prohibitive on these networks. Focusing on a dynamic latent factor model, we embed a variational Bayesian approach within an online estimation scheme in order to model activity on a network with tens of thousands of nodes.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

I am profoundly grateful to the many people who made this journey possible. First, I would like to thank my mentor and advisor, Tyler McCormick. Over the last five years, Tyler has been instrumental in my development from a student to a contributing member of the statistics community. He has always provided his unconditional support, and without him a career studying social networks would not be possible. He has impressed upon me the benefits of being able to work with scientists both in and out of the field of statistics; working with Tyler, Bailey Fosdick, Morgan Hardy, Joshua Neil, and Cole Sodja has shown me how rewarding collaboration in research can be.

I would also like to thank the rest of the faculty at the University of Chicago and the University of Washington. Michael Stein was responsible for first piquing my interest in statistics and his emphasis on critical reasoning has served me well in the years since. Debashis Mondal served as my master's thesis advisor during my last year at Chicago, and this experience ultimately led me to pursue a doctorate in the field. My committee members Adrian Dobra, Elena Erosheva, and Alan Griffith provided invaluable feedback in shaping this dissertation.

I have been fortunate to enjoy the company of many friends who have supported me along my journey. Greg Chen, Yakym Pirozhenko, and Di Ai have provided many welcome diversions since the beginning of college and hopefully for many years to come. Scott Mendelssohn, Allen Zhang, and Jason Zhou made working through long nights at the Reg much more enjoyable than they had any right to be. Nelson Auner and Zac Schaefer were good friends both in and out of the classroom during my time as a master's student.

I have had the pleasure of sharing my time at the University of Washington with some

amazing peers. David Clausen, Wayne Yang, Edward Zhao, and Alec Zimmer were some great roommates in some not so great apartments. Chris Aicher, Amrit Dhar, Corrine Jones, Eric Kernfeld, Kyle Lo, Amy Marsha, Hoiyi Ng, and Luca Weihs have continuously pushed me to become a better statistician and spent perhaps too many hours with me playing board games such as One Night Ultimate Werewolf. I look forward to their inevitable migration to the Bay Area, but in the meantime hope to continue questing together as a merry band of adventurers in D&D.

Last, but certainly not least, I would like to thank my family for their support from the very beginning. My sisters Annie, Suzanne, and Meg have always encouraged me in all endeavors and served as founts of advice. My parents Amy and Lung-Fei taught me that I could achieve anything I set my mind to and have always pushed me to challenge myself both academically and personally. I am forever grateful to have them as parents.

# DEDICATION

For my father, Lung-Fei Lee

# Chapter 1

# INTRODUCTION

Networks, or equivalently, graphs, are used to conceptualize complex systems in many fields, including the social sciences, economics, and information technology. We mathematically represent these systems in terms of graphs, a collection of nodes and edges connecting pairs of nodes (dyads). In a social network, nodes may represent individuals in a population of interest and edges may represent the relationships between these individuals. In an economic network, nodes might be used to represent economic agents (e.g. an individual, business, or government) while edges might denote economic or financial ties between these entities. In a communication network, nodes can be used to represent computers and edges can be used to represent information flow between computers.

In order to gain insight into these complex systems, network data on the relationships between nodes are often directly studied. In economic settings, researchers may be interested in identifying the most influential or well-connected agents in a network; interventions which target these agents may have a larger impact on the overall population than interventions which randomly select agents (Banerjee et al., 2013). As another example, network information can be aggregated at the node level and included as explanatory variables for node-level regression models. In economics, these models have been used to study the influence of an individual's connections on their own behavior, particularly for educational attainment (e.g. Sacerdote (2001)).

A key assumption made in these analyses is that the network of interest is accurately measured, e.g. in a network regression model knowing which individuals influence each other's behavior. We contest this assumption, noting that networks in social sciences and economics are most commonly measured via surveys and are prone to mismeasurement due

to misreporting and survey design decisions. Killworth and Bernard (1976) observes a strong level of disconnect between subjects' reported and observed behavior with other subjects. Hardy and McCasland (2019) finds low rates of reciprocity when asking about theoretically undirected relationships. Additionally, surveys may limit the number of possible connections for each individual due to concerns about survey fatigue (e.g. Oster and Thornton (2012); Cai et al. (2015)). Other empirical literature (Bell et al., 2007; Marsden, 2016) address the reliability of survey data with emphasis on the type and salience of relationships being surveyed, temporal dependence, and how links are elicited.

In Chapter 2, we demonstrate an instance of how directly using the observed network data for inference may affect results when there are discrepancies between the observed data and the network of interest. Specifically, in the context of experiments on networks, we consider the impact of a mismeasured network on estimates of treatment effects. Through simulations, we show that mismeasured connections can induce bias for existing estimators of treatment effects. In order to address this bias, we conceptualize the observed network data as a noisy manifestation of the true latent network of interest. We develop a mixture model that incorporates the uncertainty between the observed and latent networks, prove it provides consistent estimators under assumptions about the corruption mechanism, and demonstrate how this model can be used to study the spread of weather insurance knowledge among farmers in rural China.

In the rest of this thesis, we focus on an alternative source of network data: relational event data (Butts, 2008), consisting of dyadic interactions or activity measured over time. As mobile technology becomes nearly ubiquitous, relational event data is being collected at unprecedented rates and with increasingly granular temporal precision. Call detail records (CDRs), for example, provide detailed descriptions of mobile phone interactions on a national scale (Blumenstock, 2012). The widespread adoption of social media has led to a wealth of continuous-time activity data (Sadilek et al., 2012). In addition, in ecology, advances in tracking devices have resulted in an increase in animal telemetry logs documenting movement and interactions between animals (Rutz and Hays, 2009). Relational event data offer

an appealing alternative to traditional, survey-based network measures. Relational events on a network are typically passively collected using automated logging technology, as opposed to explicitly surveying a population in order to elicit a network for research purposes. Consequentially, collecting relational event data is often relatively inexpensive and less prone to data quality issues. It can also allow researchers to study larger populations of actors at higher temporal resolutions and for more extended periods (Watts, 2007).

However, the use of relational event data presents additional modelling challenges as well. Interaction data consist only of measurements *when individuals interact*, and the absence of interaction should not be taken as explicit declaration of no relationship. More generally, interaction counts between dyads should not necessarily be taken as a direct measure of the strength of the underlying dyadic relationship. Consider email records in which employee A emailed coworker B multiple times per week and emailed his/her manager C once every other week. From our perspective, both of these email patterns may indicate strong relations between employee A and the others. Although communication with the manager is relatively infrequent compared to the communication with the coworker, the relationship between A and his/her manager should be classified as strong and significant.

Again, we seek to explicitly draw a distinction between the observed network data and the underlying network of relationships. Rather than directly quantifying relationships, relational events instead serve as manifestations of these relations and provide evidence of the underlying social structure. This approach is congruent with the fundamental approach used in random graph models (e.g. Frank and Strauss (1986); Wang and Wong (1987); Hoff et al. (2002)), in that the observed network is viewed as a stochastic instantiation of an underlying process or model. However, these models often reflect the traditional survey-based sources of data they were conceptualized for and inadequately handle certain peculiarities associated with relational event data. We seek to begin bridging this divide in Chapters 3 and 4 with respect to the continuous-time nature and potentially large scale of relational event data, respectively.

Random graph models typically assume the observed network data arise in the form of

a sociomatrix, an adjacency matrix with entries indicating relational strength. Discrete-time temporal dynamics are then introduced to these models (e.g. Krivitsky and Handcock (2014); Durante and Dunson (2016); Sewell and Chen (2015)) by considering a series of these sociomatrices over time. In contrast, relational event data are measured in continuous-time, and so in Chapter 3 we propose a point-process model for inferring a network of social relations that directly models relational activity in continuous-time. We model interactions with inhomogeneous Poisson processes whose intensities are dependent on time, covariates, and the dynamic latent network. Interactions are allowed to be spurious and not inherently indicative of an underlying connection; rather, these latent relations are characterized by consistent deviations from expected, baseline behavior. We explore networks inferred by our method in the contexts of college students and barn swallows.

In Chapter 4, we instead focus primarily on the scalability challenge posed by modeling relational event data, which allow researchers to study networks of a much larger scale than survey sources. With scalability in mind, the balance between a rich and parsimonious model becomes particularly salient, and we choose to discretize the continuous-time data and model the underlying network with a dynamic version of the latent factor model (Hoff, 2005), a close analog of the latent position model (Hoff et al., 2002) designed for directed data. The latent factor model is able to capture dyadic activity while remaining relatively parsimonious by using node-specific parameters and sharing information across dyads involving the same node. A particular point of emphasis for this chapter is anomaly detection, and to this end we develop an online estimation algorithm via a variational Bayesian approach. Estimation is augmented with a case-control approximation to take advantage of the sparsity of the network and reduces computational complexity from $O(N^2)$ to $O(E)$, where $N$ is the number of nodes and $E$ is the number of edges. We run our algorithm on network event records collected from an enterprise network of over 25,000 computers in order to identify potential cybersecurity attacks.

# Chapter 2

# ESTIMATING SPILLOVERS USING IMPRECISELY MEASURED NETWORKS

Joint work with Morgan Hardy, Rachel M. Heath, and Tyler H. McCormick.

## *2.1 Introduction*

Interactions between peers are of interest in experiments in many economic settings, such as health (Oster and Thornton, 2011; Godlonton and Thornton, 2012), education (Angelucci et al., 2010; Duflo et al., 2011), job search (Magruder, 2010; Wang, 2013; Heath, 2018), personal finance (Bursztyn et al., 2014), agriculture (Cai et al., 2015; BenYishay and Mobarak, 2018; Beaman et al., 2018; Vasilaky and Leonard, 2018), and microenterprises (Hardy and McCasland, 2019). In these experiments, in addition to their own treatment assignments, individuals may also be influenced by the treatment assignments of their peers. These treatment spillovers are often of direct scientific interest, representing important behavior to account for when making policy decisions. Moreover, even when treatment spillovers to peers are not of direct interest, the possibility of treatment spillovers to the control group violates the stable unit treatment value assumption (SUTVA) needed to identify causal treatment effects (Rubin, 1974). In both cases, knowing the group of peers who are potentially affected by a treatment allows researchers to accurately estimate peer effects and assess potential SUTVA violations.

Despite the challenges associated with accurately measuring the network of treatment interference, existing methods for estimating treatment effects assume the observed network data perfectly captures this network of interest. In this chapter, we consider the setting where the observed network instead represents a mismeasured version of the true network, allowing

for both unreported spillover pathways and misreported links over which no spillovers could occur. Focusing on treatment effect estimators under the local network exposure approach (Ugander et al., 2013; Aronow and Samii, 2017), which is a potential outcomes framework that defines "treatment exposure conditions" based on treatment assignment of each subject and their direct connections, we first show missing links and misreported links in the network can cause mismeasured treatment exposure conditions and hence biased estimators. In order to recover unbiased estimators, we develop a class of mixture models that accounts for the uncertainty of the latent true exposure conditions and discuss parameter estimation using the Expectation-Maximization (EM) algorithm. These models rely on parametric assumptions about the distribution of missing links conditional on the observed network data as well as parametric assumptions on the behavior of outcomes within each treatment exposure condition[1]. Under a linear regression model for the latter assumptions, we prove the mixture model is identifiable and the maximum likelihood estimator from the EM algorithm is consistent.

We evaluate our method with both simulations and replication of an existing study. We simulate experiments on networks of Indian households from 75 villages (Banerjee et al., 2013). We are able to recover accurate estimates of direct and indirect treatment effects when commonly used Horvitz-Thompson estimators based on weighted averages of outcomes by group fail. Finally, we implement our method using networks data from a randomized evaluation of insurance information sessions with rural farmers in China (Cai et al., 2015). We find that our method produces more consistent estimates of direct and indirect treatment effects, across various choices of network measures, than naive treatment effects estimates that assume the network is measured perfectly.

---

[1]In the context of experiments, mixture modeling has previously been used under the potential outcomes framework to address subject compliance (Sobel and Muthén, 2012). Subjects are classified into various conditions based on their behavior with respect to treatment assignment (e.g. never takes treatment, complies with treatment, always takes treatment), with the goal of measuring a treatment effect solely for complier subjects. However, this classification is inherently unknown since the behavior of each subject is only observed under a single treatment assignment, and thus estimation proceeds by jointly modeling the uncertainty over these compliance conditions with the treatment outcome under each compliance condition.

Our results are relevant to many experimental contexts where subject's behavior or outcome may be influenced by other subjects' treatment assignment in addition to their own. A common approach in these cases is to randomize treatment at a geographic or organizational level that plausibly contains each treated individual's network of potential spillovers, such as a village (in isolated, rural settings), and then compare treated individuals to "pure controls" in non-treated units. However, even if this is possible, comparing treated to control subjects still confounds treatment effects and spillovers on these treated subjects.[2] Moreover, in other cases, a pure control is not feasible, because the experiment must be implemented within a single firm (Bandiera et al., 2009; Bloom et al., 2014; Adhvaryu et al., 2019) or market (Conlon and Mortimer, 2013), or it is not possible to leave a large enough buffer between treated and control areas to render spillovers unimportant. Potential SUTVA violations could then introduce both upward and downward bias in direct treatment effect estimates.[3]

The local network exposure approach (Aronow and Samii, 2017; Ugander et al., 2013) that we use contrasts with linear-in-means models (Manski, 1993; Bramoullé et al., 2009) in its assumed avenues of treatment interference. Local network exposure models assume the avenues of interference for each subject are limited to the treatment assignments of other subjects in their direct (first-order) network. Discrete treatment "exposure conditions" are defined based on a subject's and their connections' treatment assignments, and are used within Rubin's potential outcome framework as a set of potential treatment conditions (as opposed to using the levels of treatment). On the other hand, linear-in-means models (Manski, 1993; Bramoullé et al., 2009) hypothesize indirect treatment effects manifest as

[2]An exception would be if the treated individuals are a small enough fraction of treated units that they are unlikely to know treated subjects. Comparing treatment to control individuals would then identify the average direct treatment effect by construction. However, this would likely require a large enough number of units to be impractical or prohibitively expensive in many settings. Treatment effects in such contexts are also not particularly informative about the results from scaling up a treatment to an entire population.

[3]The reduction in exposure to disease from directly treated school children in Kenya may indirectly improve the health outcomes of school children who did not directly receive the treatment, biasing downward naively estimated benefits of deworming pills (Miguel and Kremer, 2004). In contrast, increased police patrolling on the streets of Bogota, Colombia, may merely push crime "around the corner", biasing upward the estimated impact on crime rates (Blattman et al., 2017).

a linear relationship between a subject's outcome and the average treatment and average outcome of that subject's peers. The dependence between outcomes of connected subjects allows for the a subject's outcome to be influenced by any other subject to which they are directly or indirectly connected to in the social network, with the amount of influence being modulated by their distance in the network. While we use a local network exposure model that allows us to focus on the effect of exposure to at least one treated subject, in Section 2.2.2, we explain how this approach can allow for similar dynamics as a linear-in-means model if we increase the number of indirect treatment bins assumed to influence an individual.

Our approach is related to a growing literature in economics, political science, sociology, and statistics on network sampling. One common setting is assumes that the researcher can perfectly observe a fraction of the total network. For example, Chandrasekhar and Lewis (2011) shows how egocentrically sampled network data can be used to predict the "full" network in a process they term graphical reconstruction[4]. By contrast, we study a setting in which all potential links are measured, but may contain some error. As in Handcock and Gile (2010) and Newman (2018), we relate the observed and latent network via a probabilistic model and, given a set of model parameters, construct a distribution over the true network conditional on the observed graph[5].

This chapter proceeds as follows. In the next paragraph, we introduce notation that we will use throughout. Then, in Section 2.2 we discuss existing methods for estimating direct and indirect treatment effects. In Section 2.3, we derive formulas for the bias in Horvath-Thomson estimators based on weighted averages when networks are measured with error. In Section 2.4 we propose a mixture model to estimate treatment effects that can account for latent ties between subjects. We discuss when this model is identified, how to estimate model parameters and treatment effects, and examine model performance using simulations.

---

[4]See Williams (2016) and Griffith (2017) for sample applications.

[5]In our setting, by contrast, even a full graph cannot be used to train probabilistic models, because of the potential for error on every link (and non-link). This creates an inability to learn the parameters of the mismeasurement process. For example, the observed network data does not inform the proportion of true links missing from the observed graph.

We apply our methodology to the agricultural setting of Cai et al. (2015) in Section 2.5, and conclude in Section 2.6 with a discussion.

We now introduce some basic notation that we will use throughout the rest of the chapter. Let $i \in \{1, 2, \cdots, N\}$ index the subjects in the study, with corresponding observed outcomes $y_i$, which we denote as $\mathbf{y}$. For simplicity, suppose treatment is binary with levels "treatment" (1) and "control" (0), and the treatment assignment mechanism is random and explicitly known. Denote the vector of treatment assignment with $\mathbf{t} \in \{0, 1\}^N$, in which the treatment of individual $i$ is $t_i$. Suppose the true influence network $G$ is directed and binary, with the edge $i \to j$, representing individual $i$'s influence on individual $j$, encoded by $G_{ij} = 1$. Let $G_j$ denote the $j$th column of $G$, indicating the influencers of individual $j$, so $\mathbf{1}'G_j$ is the the number of influencers or in-degree of $j$.[6] For now, let us assume $G$ is observed. Finally, let $\overline{G_j}$ denote the $j$th column of G normalized to sum to 1 ($\mathbf{1}'\overline{G_j} = 1$) unless $\mathbf{1}'G_j = 0$, in which case $\overline{G_j} = G_j = \mathbf{0}$.

## *2.2 Measuring network spillovers in experiments*

### *2.2.1 Local network exposure model*

Aronow and Samii (2017) and Ugander et al. (2013) propose estimators for average direct and indirect treatment effects by building on the Rubin causal model (Rubin, 1974). In the context of experiments, each subject has a set of "potential outcomes" $(Y_i(t_i = 0), Y_i(t_i = 1))$ corresponding to the possible outcomes under each treatment (or none). The inference task is to estimate the average treatment effect, defined to be the difference between the average outcome of the population if the entire population was treated and the average outcome if the entire population was in the control:

$$ATE(1, 0) = \frac{1}{N} \sum_{i=1}^{N} \left[ Y_i(t_i = 1) - Y_i(t_i = 0) \right]. \tag{2.1}$$

[6] Analogously, $G_j \cdot \mathbf{1}$ is the number of people that individual $j$ influences, or the out-degree.

This quantity is not observed since we cannot observe the full set of potential outcomes for each subject, but assuming completely random assignment can be estimated by the difference in sample means:

$$\widehat{ATE(1,0)} = \frac{1}{N_1}\sum_{i=1}^{N} y_i \mathbf{1}[t_i = 1] - \frac{1}{N_0}\sum_{i=1}^{N} y_i \mathbf{1}[t_i = 0], \tag{2.2}$$

where $N_k$ is the number of subjects in treatment $k$. A crucial assumption in the Rubin causal model is SUTVA, which states than a subject's potential outcomes are unaffected by the treatments of other subjects. In experiments on networks, SUTVA is violated if the treatments of peers influence the outcomes for an individual.

Aronow and Samii (2017) considers a violation of SUTVA by allowing for individuals to be systematically affected by the treatment assignments of their peers. By making assumptions that restrict the nature of these influences, they induce mappings of the treatment vector $\mathbf{t}$ to distinct "exposure conditions", or what Manski (2013) terms "effective treatments." In a simple instance of their framework, which we borrow for our model in Section 2.4, individuals are affected by whether or not any of their influencers in $G$ are treated, inducing assignments into one of four exposure conditions, corresponding to levels of direct and indirect exposure to treatment:

$$C_i \equiv C_i(t, G_i) = \begin{cases} c_{00} & \text{(No Exposure)}: \quad t_i = 0 \text{ and } \mathbf{t}'G_i = 0 \\ c_{01} & \text{(Indirect Exposure)}: \quad t_i = 0 \text{ and } \mathbf{t}'G_i > 0 \\ c_{10} & \text{(Direct Exposure)}: \quad t_i = 1 \text{ and } \mathbf{t}'G_i = 0 \\ c_{11} & \text{(Full Exposure)}: \quad t_i = 1 \text{ and } \mathbf{t}'G_i > 0. \end{cases} \tag{2.3}$$

In this model, both direct and indirect effects are taken to be binary, with an individual being indirectly exposed to treatment if one or more of their influencers are (directly) treated. Each subject $i$ would have four potential outcomes $(Y_i(C_i = c_{00}), Y_i(C_i = c_{01}), Y_i(C_i = c_{10}), Y_i(C_i = c_{11}))$, one for each exposure condition. Note this setup assumes that the number of connections treated does not have an effect beyond the presence or absence of at least

one, and an individual can only be influenced by a first-order connection in the network.

The choice of indirect exposure can be related to diffusion models of information and disease in which "contagion" can occur given a single source of exposure (Centola and Macy, 2007), also called simple contagion models. In a Bayesian learning framework, these models would be relevant in cases where individuals do not have strong priors (so that the first piece of information they receive is the most important) and where the information is non-rival and relatively costless to pass along (so that a treated network member would be very likely to pass on information). Moreover, rational individuals in a Bayesian learning context infer that information shared by multiple network members is likely come from a common source, so additional information will be less valuable.[7]. By contrast, in settings in which individuals have a stronger prior – which is different from the information they receive – or information is costly to pass along, the number of treated peers matters; these settings are sometimes called complex contagion models. For instance, an individual adopts a technology if a fraction of her network that is above some threshold adopts the technology (Granovetter, 1978; Acemoglu et al., 2011; Beaman et al., 2018). In such cases, the assumption that only the first treated peer matters can be relaxed by adding additional exposure conditions that correspond to the appropriate model of peer influence in a given context.[8] Similarly, the assumption that only first-order links matter could be relaxed by adding exposure conditions corresponding to second-order exposure.

The primary quantities of interest would then be given by average treatment effects akin to equation (2.1):

$$ATE(c,d) = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i(C_i = c) - Y_i(C_i = d)\right]. \tag{2.4}$$

The average direct treatment effect would be given by $ATE(c_{10}, c_{00})$, while the average

[7]Alternatively, other social learning models assume that individuals are boundedly rational and do not infer that information shared by multiple network members likely comes from a common source (DeGroot, 1974; Banerjee et al., 2019).

[8]For instance, Beaman et al. (2018) find evidence in favor of a threshold model in which at least two treated peers is necessary for adoption of a new technology.

indirect treatment effect when not directly treated would be $ATE(c_{01}, c_{00})$. Estimating these quantities is equivalent to estimating the mean outcomes of the entire population under each exposure condition:

$$\mu_c = \frac{1}{N} \sum_{i=1}^{N} Y_i(C_i = c), \tag{2.5}$$

so we focus on the latter for this section and the next, with the additional assumption that each subject is assigned to treatment with some constant and known probability independently of other subjects. In contrast to the case when the SUTVA assumption is satisfied, we cannot estimate these means using just their sample counterparts. Under this random treatment assignment mechanism, the probability that each subject is placed in each exposure condition is (generally) not equal. While the probability of direct treatment is constant across all subjects, variability in the in-degrees of individuals causes variation in the probabilities of assignment to each indirect treatment level and thus each exposure condition. Namely, individuals with high in-degree are more likely to be indirectly exposed to the treatment since they have more influencers who potentially may be treated. This selection bias could affect the mean estimates if there is heterogeneity in the outcomes within exposure conditions associated with in-degree. Horvitz-Thompson estimators use inverse probability weighting in order to take varying exposure probabilities into account to produce unbiased estimators of these mean outcomes:

$$\widehat{\mu}_{c,HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \cdot \mathbf{1}_{[C_i=c]}}{P(C_i = c)}. \tag{2.6}$$

This estimator is equivalent to the sample mean if the probability of assignment to an exposure condition is constant among subjects. These estimators are unbiased regardless of the form of the heterogeneity between the outcomes and network degrees.[9] Note if certain subjects have zero probability of being placed in certain exposure conditions, e.g. when a

[9]However, they can have high variance when the exposure conditions are highly imbalanced on in-degree. This would arise when the probabilities $P(C_i = c)$ are small for some $i$, yielding large weights $\frac{1}{P(C_i=c)}$. This suggests potential efficiency gains from stratifying on degree.

subject has no influencers, estimation must be restricted to the sub-population of individuals with non-zero probability of being placed in every condition.

Explicitly modeling the relationship between potential outcomes and network degrees can result in lower variance estimators at the cost of additional assumptions about the validity of these relationships. For example, suppose we believe that for each exposure condition $c$, the relationship between the in-degree ($\mathbf{1}'G_i$) and the potential outcome $Y_i(C_i = c)$ can be modeled with

$$Y_i(C_i = c) \sim f\left(\cdot; \theta_c, \mathbf{1}'G_i\right), \tag{2.7}$$

where $\mathbf{1}'G_i$ is the in-degree of individual $i$ and $\theta_c$ are model parameters. Assuming this model accurately characterizes the relationship between the potential outcomes and in-degrees, the distribution of potential outcomes is conditionally independent of the exposure assignment (induced by the treatment assignment) vector given the in-degrees of the subjects, such that the exposure assignment mechanism can be "ignored" during inference of the means (Rubin, 1974). The estimate of the mean outcome under exposure condition $c$ would then be given by

$$\widehat{\mu}_{c,R} = \frac{1}{N}\sum_{i=1}^{N} \widehat{Y}_i(C_i = c) = \frac{1}{N}\sum_{i=1}^{N} E_{f\left(\cdot;\widehat{\theta}_c,\mathbf{1}'G_i\right)}[y_i], \tag{2.8}$$

provided an estimate of model parameters $\widehat{\theta}_c$. Parametric models $f\left(\cdot; \theta_c, \mathbf{1}'G_i\right)$ of the outcomes under condition $c$ and in-degree $d$ are necessary for likelihood-based approaches to estimation and are used in the model we propose in Section 2.4. A common model familiar to many economists is

$$f\left(\cdot; \theta_c \equiv (\alpha_c, \beta_c, \sigma^2), \mathbf{1}'G_i\right) = \mathrm{N}\left(\cdot; \alpha_c + \beta_c \mathbf{1}'G_i, \sigma^2\right), \tag{2.9}$$

which corresponds to a linear model with different intercepts and slopes for each exposure condition (but common variance). In this case, the estimates of mean outcome would be given by $\widehat{\mu}_{c,R} = \widehat{\alpha}_c + \widehat{\beta}_c \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}'G_i$.

### *2.2.2 Relation to the linear-in-means model*

Although we focus on the local network exposure model in the subsequent sections, let us first compare it to another popular approach for accounting for and measuring treatment spillovers: Manski linear-in-means models (Manski, 1993; Bramoullé et al., 2009). In the context of experiments without additional covariates, a basic form of these models would be

$$y_i = \alpha + \beta y' \overline{G_i} + \gamma t_i + \delta t' \overline{G_i} + \epsilon_i, \tag{2.10}$$

In addition to allowing for "endogenous effects" ($\beta \neq 0$), which would account for second- and higher-order indirect treatment effects (e.g. how an individual is affected by their peers' peers' outcomes), the linear-in-means-model differs from the described local network exposure model by placing a different set of assumptions on the mechanism of indirect treatment effects. Rather than partitioning indirect effects into varying magnitudes based on indirect exposure conditions, the linear-in-means model assumes the size of indirect effects vary linearly with the proportion of peers treated. Note a similar assumption could be used with the local network exposure setting by characterizing indirect exposure with proportions of peers treated instead of the presence of any peers treated, albeit these proportions would have to be arranged into discrete bins. Similarly, we could introduce a non-linear dependence on the number of treated peers in the linear-in-means framework by adding multiple indicator variables to Equation 2.10. In both cases, introducing more discrete potential outcome categories will likely lead to small cell counts in practice. This observation highlights the importance of the linearity assumption in the linear-in-means model.

## *2.3 Characterizing the bias in local network exposure model under mismeasurement*

In this section we derive the bias in Horvitz-Thompson estimators (2.6) if using a mismeasured network, $\widetilde{G}$, to estimate exposure conditions instead of the true network $G$. We allow $\widetilde{G}$ to be mismeasured such that there are either links present in $\widetilde{G}$ that are not in $G$ or

vice-versa. Suppose our treatment assignment mechanism $\mathbf{t}$ is constructed such that each subject $i$ has positive probability of being placed in treatment and positive probability of being placed in control. We can break the impact of using $\widetilde{G}$ in estimation into two distinct factors. First, note that the Horvitz-Thompson estimator can only be used for subjects with non-zero probability of being placed in each exposure condition. Namely, subjects with zero in-degree must be excluded, reflecting the idea that a potential outcome under indirect exposure only makes sense if the subject could be indirectly exposed to treatment under some hypothetical treatment assignment. When we observe a mismeasured version of the network, we may not be able to accurately identify which subjects should be excluded. Certain individuals who have positive in-degree in $G$ may be observed to have zero in-degree in $\widetilde{G}$ and thus would be incorrectly excluded for estimation. At the same time, certain individuals with $\mathbf{1}'G_i = 0$ may be observed to have positive in-degree and thus be included during estimation. If either of these situations arose, our estimated average outcomes would then represent a different subpopulation than the true population of subjects with non-zero in-degree.

Second, even if we are able to accurately identify all subjects with non-zero in-degree, bias in mean estimates may be induced by distorted observed exposure conditions. Subjects who are in truth indirectly exposed to treatment would not be observed to be indirectly exposed if all connections to influencers who are treated are unobserved (and no false links to other treated individuals are observed). Similarly, subjects not indirectly exposed to treatment may be falsely observed to be indirectly exposed. The mismeasured exposure conditions are able to correctly identify the level of direct treatment for each subject but not necessarily the level of indirect treatment. Mathematically, observing $\widetilde{C}_i \equiv C_i(\mathbf{t}, \widetilde{G}_i) = c_{kl}$ for any $k, l \in \{0, 1\}$ may correspond to either $C_i(\mathbf{t}, G_i) = c_{k0}$ or $C_i(\mathbf{t}, G_i) = c_{k1}$. Recall that in this notation the first subscript denotes the direct treatment condition (whether $i$ is directly treated or not) and the second subscript denotes the indirect treatment (whether at least one member of $i$'s network was treated). The Horvitz-Thompson estimators for each treatment exposure condition $c$ under the mismeasured network $\widetilde{G}$ are given by

$$\widehat{\mu}_{c,HT,\widetilde{G}} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i \cdot \mathbf{1}_{[\widetilde{C}_i=c]}}{P(\widetilde{C}_i = c)}. \tag{2.11}$$

where observed $y_i = \sum_c Y_i(C_i = c)\mathbf{1}_{[C_i=c]}$ is dependent on the true exposure condition and the probabilities are taken over possible treatment assignments $\mathbf{t}$. Holding the observed and true networks fixed and taking the expectation of the estimators $\widehat{\mu}_{c,HT,\widetilde{G}}$ over the possible treatment assignments $\mathbf{t}$ we have:

$$E\left[\widehat{\mu}_{kl,HT,\widetilde{G}}\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{E\left[y_i \cdot \mathbf{1}_{[\widetilde{C}_i=c_{kl}]}\right]}{P(\widetilde{C}_i = c_{kl})} \tag{2.12}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_c Y_i(C_i = c)P\left(\widetilde{C}_i = c_{kl}, C_i = c\right)}{P(\widetilde{C}_i = c_{kl})} \tag{2.13}$$

$$= \frac{1}{N}\sum_{i=1}^{N} Y_i(C_i = c_{k0})P\left(C_i = c_{k0}|\widetilde{C}_i = c_{kl}\right) + Y_i(C_i = c_{k1})P\left(C_i = c_{k1}|\widetilde{C}_i = c_{kl}\right). \tag{2.14}$$

We find the mean estimate of the $c_{kl}$ conditioned on $\widehat{\mu}_{kl,HT,\widetilde{G}}$ under the mismeasured network $\widetilde{G}$ tends to lie between the mean outcomes under the two exposure conditions corresponding to the same level of direct treatment: $\mu_{k0}$ and $\mu_{k1}$. The bias will be greater, the greater is the probability of mismeasurement ($P(C_i = c_{k0}|\widetilde{C}_i = c_{kl})$ and $P(C_i = c_{k1}|\widetilde{C}_i = c_{kl})$ are far from 1) and the greater is the difference in outcomes between those who are actually indirectly treated versus not ($Y_i(C_i = c_{k0})$ is far from $Y_i(C_i = c_{k1})$. In section 2.4.4, we will perform a simulation study that investigates the level of bias in these Horvath-Thompson estimators and our proposed EM estimates at different rates of unreported true links and falsely observed links.

## *2.4 Latent Variable Model for Network Spillovers*

In this section we propose a latent variable approach to estimating average treatment effects when the network observed is a noisy representation of the true network of interest. We

assume that each true edge $G_{ij} = 1$ is not observed ($\widetilde{G}_{ij} = 0$) with probability $p$, non edges $G_{ij} = 0$ are observed with probability $q$, and edges are observed/not observed independently of one another. These corruption mechanisms assume the observed edges are a random subset of the true edges and the false edges are a random subset of the non-edges. In real economic networks, particularly those with high node heterogeneity, this might be an tenuous assumption, but we can relax this assumption to allow adding/subtracting edges to depend on observed covariates, and do so in the empirical application in section 2.5.

### *2.4.1 Latent Variable Model*

Suppose the true network of interest $G$ is unobserved and we only observe a mismeasured network $\widetilde{G}$. Furthermore, assume the effects of treatment can be characterized with the four exposure conditions defined in (2.3). For individual $i$, we observe mismeasured exposure condition $C_i(t, \widetilde{G}_i)$ and in-degree $\mathbf{1}'\widetilde{G}_i$. The statistical problem is then to model the relationship between these mismeasured statistics and their true, latent, counterparts. Given a distribution over the true exposure condition $C_i(t, G_i)$ and in-degree $\mathbf{1}'G_i$, we can use models like those in equations (2.7) and (2.8) to estimate mean outcomes for each exposure category. For notational simplicity, let $\widetilde{C}$ represent the vector of mismeasured exposure conditions, $\mathbf{1}'\widetilde{G}$ the vector of mismeasured in-degree, and $C$ and $\mathbf{1}'G$ the corresponding latent terms.

Consider subject $i$, who has exposure condition $C_i(t, G_i) \equiv c_{kl}$, degree $\mathbf{1}'G_i \equiv d$, and $t'G_i \equiv d_t$ connections with treated subjects, but for whom we observe exposure condition $C_i(t, \widetilde{G}_i) \equiv c_{\tilde{k}\tilde{l}}$, degree $\mathbf{1}'\widetilde{G}_i \equiv \tilde{d}$, and $t'\widetilde{G}_i \equiv \tilde{d}_t$ connections with treated individuals instead. Holding the treatment assignments to be fixed, we can separately model the number of connections to treated subjects $d_t$ and the number of connections to not-treated subjects $d - d_t$, from which we can derive the induced exposure conditions. Note this procedure works for any indirect exposure conditions entirely characterized by the number of treated connections and the number of total connections (e.g. ratio of treated connections) and not just (2.3). Following Bayes' rule and noting we observe $\tilde{d}_t$ treated connections when $x$ of the $d_t$ actual treated connections are dropped and another $\tilde{d}_t - d_t + x$ false connections to

treated individuals are observed,

$$P(t'G_i = d_t | t, t'\widetilde{G}_i = \tilde{d}_t; p, q) \propto P(t'\widetilde{G}_i = \tilde{d}_t | t, t'G_i = d_t; p, q) P(t'G_i = d_t) \tag{2.15}$$

$$\propto \sum_{x=0}^{d_t} \text{Bin}(x; d_t, p) \text{Bin}(\tilde{d}_t - d_t + x; \mathbf{1}'t - t_i - d_t, q) P(t'G_i = d_t) \tag{2.16}$$

where $\text{Bin}(x; n, p)$ is the probability of $x$ successes from a binomial distribution with $n$ attempts and success probability $p$. Similarly for connections for non-treated subjects,

$$\begin{aligned} P((1-t)'G_i = d_{nt} | t, (1-t)'\widetilde{G}_i = \tilde{d}_{nt}; p, q) \propto & \sum_{x=0}^{d_{nt}} \text{Bin}(x; d_{nt}, p) \\ & \times \text{Bin}(\tilde{d}_{nt} - d_{nt} + x; N - 1 - \mathbf{1}'t + t_i - d_{nt}, q) \\ & \times P((1-t)'G_i = d_{nt}) \end{aligned} \tag{2.17}$$

Both sets of equation require a (prior) model over the number of true connections to treated and un-treated subjects. Assuming no additional information about the structure of the true network, one of the most simplistic models would be to model the true graph as an Erdos-Renyi graph, where the probability of a link between any given edges is constant, leading to independence across edges. Under this model, the number of connections to treated/un-treated subjects could be modeled with binomial distributions. However, in many real-world networks we find that the degree distribution demonstrates extra-binomial variation, where differences in degree arise not just from random variation in link formation but also from differences in the propensity to form links. Thus in the following sections we prefer to use a beta-binomial model. With a beta-binomial distribution, we can think of each degree as being sampled from a binomial distribution $d \sim \text{Binom}(N-1, p)$, where $p$ is independently sampled from a Beta distribution $p \sim \text{Beta}(\mu, \rho)$, where we paramaterize the beta-binomial distributions in terms of an average probability of success $\mu$ and an overdispersion parameter

$\rho$. The variance of this beta-binomial would be given by $(N-1)\mu(1-\mu)\left(1+(N-2)\rho\right)$, compared to $(N-1)\mu(1-\mu)$ for a binomial distribution with parameter $\mu$. We leave the these parameters to be chosen on a application-by-application basis, noting that the choice of these parameters are more influential when there is high mismeasurement in the network and hence higher uncertainty over the true degrees [10].

Using the above equations, we can express the relationship between the true exposure condition and degree and their observed counterparts:

$$\tau_i(c_{kl}, d; p, q) \equiv P(C_i(t, G_i) = c_{kl}, \mathbf{1}'G_i = d | \widetilde{G}_i, t; p, q) \tag{2.18}$$

$$= \begin{cases} P(t'G_i = 0, \mathbf{1}'G_i = d | t, \widetilde{G}_i; p, q), & l = 0 \\ P(t'G_i > 0, \mathbf{1}'G_i = d | t, \widetilde{G}_i; p, q), & l = 1 \end{cases} \tag{2.19}$$

$$= \begin{cases} P(t'G_i = 0 | t, \widetilde{G}_i; p, q) P((1-t)'G_i = d | t, \widetilde{G}_i; p, q), & l = 0 \\ \sum_{d_t=1}^{d} P(t'G_i = d_t | t, \widetilde{G}_i; p, q) P((1-t)'G_i = d - d_t | t, \widetilde{G}_i; p, q), & l = 1 \end{cases} \tag{2.20}$$

Equation (2.20) defines a distribution over the true, unobserved exposure condition and in-degree, conditional on the treatment vector and the number of observed treated and non-treated connections for individual $i$. When coupled with a parametric model $f(\cdot; \theta_c, d)$ (see 2.7) for the potential outcomes under each (true) exposure condition $c$ and in-degree $d$, we can model the observed outcome $y_i$ as arising from a mixture of the $f(\cdot; \theta_c, d)$ with weights corresponding to the probabilities $\tau_i(c_{kl}, d; p, q)$ over the unobserved quantities (namely, true treatment status $c_{kl}$ and degree $d$).[11] The log-likelihood of the parameters

[10] Via simulations, we find that a good choice of $\mu$, which governs the overall density of the true network, is more important for our model to recover unbiased estimates.

[11] One downside of the Horvitz-Thompson estimator (2.6) is that it does not model individual potential outcomes and thus is less amenable to likelihood-based approaches.

$\Theta \equiv \{\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}, p, q\}$ given $y$ is

$$l(\Theta) = \sum_{i=1}^{N} \log \left[ \sum_{c} \sum_{d=0}^{N-1} \tau_i(c, d; p, q) f\left(y_i; \theta_c, d\right) \right]. \tag{2.21}$$

This is a mixture model in the sense that the likelihood contribution of each subject is the average of her outcome under each exposure condition, weighted by the probability of being in each exposure condition given observed data. Estimation of the parameters $\Theta$ are provided using maximum likelihood estimation via the Expectation-Maximization algorithm, details of which are provided in Section 2.4.3. Note that likelihood estimation is only justified if the observed outcomes are representative of the potential outcomes under each exposure condition, conditional on the true in-degrees. That is only the in-degree can determine indirect exposure to treatment, as in the case of a random treatment mechanism[12].

Provided an estimate of the model parameters $\widehat{\Theta}$, estimating the mean outcome under exposure condition $c$ (recall equation (2.5)) is straightforward and given by the expectation of the potential outcome under exposure $c$ for each subject averaged across the population. We estimate $\mu_c$ with the following plug-in estimator:

$$\widehat{\mu}_c = \frac{1}{N} \sum_{i=1}^{N} E\left[ Y_i(C_i = c) | y, t, \widetilde{G} \right] \tag{2.22}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{d} P(\mathbf{1}'G = d | y, t, \tilde{G}) E_{f\left(\cdot; \widehat{\theta}_c, d\right)}[y_i] \tag{2.23}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{d} \left( \sum_{c'} \tau_i(c', d; \widehat{p}, \widehat{q}) \right) E_{f\left(\cdot; \widehat{\theta}_c, d\right)}[y_i]. \tag{2.24}$$

### *2.4.2 Identification*

Before we discuss estimation strategies for our mixture model (2.21), we will (partially) characterize the conditions under which this model is identifiable. Without model identifiability,

[12]Stratified sampling based on known covariates could also be addressed by directly introducing these covariates into the model.

estimation may be unstable and parameters estimates uninterpretable. In this section, we assume $f(\cdot; \theta_c, d)$ arise from a common univariate family of distributions parameterized by $\eta \equiv \eta(\theta_c, d)$.

In general, mixture models are trivially unidentifiable since relabeling components yields different parameterizations of a model with the same marginal distribution (see Chapter 1.5 of McLachlan and Basford (1988)). In our case, for example, one could relabel direct treatment are indirect treatment and vice versa. This identifiability issue is of particular concern in our setting, where the labeling of the components is inherently meaningful; for example, being unable to disentangle clusters corresponding to no treatment and indirect treatment would leave us unable to estimate the direction of any indirect treatment effect. We are able to leverage the structure from our mismeasurement model and the linear relationships between mixture components with the same exposure condition to prevent such relabeling from occurring.

Following Frühwirth-Schnatter (2006), we use "generic identifiability" to refer to identifiability problems not solved by permuting component labels. Generic identifiability holds for mixtures of Gaussians and many other univariate continuous distributions, with the major exceptions being the binomial and uniform distributions. For the binomial distribution, generic identifiability holds if a sufficient number of trials/observations per subject are observed, dependent on the number of components. See Frühwirth-Schnatter (2006) for a review of generic identifiability issues.

Note that the fact that the model is not identified for binary outcomes means that it cannot be directly applied to settings with a single, one-time measures of technology adoption. While this is a limitation of our method given that technology adoption is an important outcome in the literature on networks, it can be applied to other measures of adoption such as input usage (Conley and Udry, 2010), or determinants of adoption such as knowledge about the new technology (as in the example from Cai et al. (2015) in section 2.5, or Beaman et al. (2018)).

Unfortunately, generic identifiability of the mixture model (2.21) does not directly follow

from the generic identifiability of the family $f$, as Hennig (2000) showed in the case of mixtures of linear regression models. For example, in a mixture of simple linear regressions with two distinct covariate values $0, 1$ and common variance $\sigma^2$, an equal mixture of $f(x) = x$ and $f(x) = 1 - x$ yields the same model as a equal mixture of $f(x) = 0$ and $f(x) = 1$. Observations from a third covariate value would yield generic identifiability. While not immediately applicable to our class of models since in-degree (our covariate) is also latent, Hennig (2000) and Grün and Leisch (2007) define conditions under which mixtures of linear and generalized linear models are generically identifiable.

Next, we explicitly prove the identifiability of our mixture model under the regression model (2.9) for $f$. Results are readily generalizable to other $f$ that arise from generically identifiable families provided that distinct values of $d$ would allow for the identification of our model parameters $\theta_c$ from the distribution parameters $\eta(\theta_c, d)$.

**Proposition 1.** *Let $f$ be defined as in (2.9) and $\tau_i$ as in (2.20). Assume $p, q, p', q' < 1$*[13] *and that indirect exposure has some effect (i.e. $\theta_{00} \neq \theta_{01}$ and $\theta_{10} \neq \theta_{11}$). Then*

$$\sum_c \sum_{d=0}^{N-1} \tau_i(c, d; p, q) f\left(y_i; \theta_c, d\right) = \sum_c \sum_{d=0}^{N-1} \tau_i(c, d; p', q') f(y_i; \theta'_c, d) \tag{2.25}$$

*for all given $t'\widetilde{G}_i = \tilde{d}_t$ and $(\mathbf{1} - t)'\widetilde{G}_i = \tilde{d}_{nt}$ implies $\{\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}\} = \{\theta'_{00}, \theta'_{01}, \theta'_{10}, \theta'_{11}\}$ as long as there exists two distinct $d$ such that we have subjects under each direct treatment status with observed degree $\tilde{d}_t + \tilde{d}_{nt} = d$, and, of these subjects, some have treated connections $\tilde{d}_t > 0$ while others do not, with $\tilde{d}_t = 0$.*

*Proof.* See appendix 2.8. □

[13] Both of these edge cases are relatively uninteresting, as when $p = 1$ all true edges are not observed and when $q = 1$ all non-edges are falsely observed.

### *2.4.3 Estimation*

Maximizing the log-likelihood (2.21) with respect to the parameters $\Theta$ cannot be done in closed form due to the summations inside the logarithmic terms. However, if we had directly observed the latent variables $\{C, \mathbf{1}'G\}$, the log-likelihood of the parameters $\Theta$ given $y$, $C$ and $\mathbf{1}'G$ would be given by

$$l(\Theta) = \sum_{i=1}^{N} \log \left[\tau_i(c, d; p, q) f\left(y_i; \theta_c, d\right)\right]. \tag{2.26}$$

This would be substantially easier to work with, due to the lack of summation inside the logarithmic terms. Essentially, estimation would entail four regressions, for each exposure condition. The EM algorithm (Dempster et al., 1977) is a well-established technique for maximum likelihood estimation in the presence of latent variables that leverages this disparity between the two log-likelihood expressions. Given some set of initial parameter values $\widehat{\Theta}^0$, the algorithm alternates between estimating posterior distribution of latent variables for each subject given the current parameter values (E-step) and updating the parameter values given these posterior probabilities (M-step). Explicitly working with the latent variables in the M-step yields simpler maximization problems. Each iteration of the algorithm increases the log-likelihood, leading to a local optimum, and the algorithm is run from multiple initialization values in order to maximize the chances of finding a global optimum.

Suppose at iteration $t$ we have parameter estimates $\widehat{\Theta}^{(t)}$. In the E-step, we compute the posterior probabilities over the latent variables using the current parameter estimates. These probabilities, or "responsibilities," are given by

$$\widehat{\gamma}_{icd}^{(t+1)} \equiv P\left(C_i(t, G_i) = c, \mathbf{1}'G_i = d | y_i, t, \widetilde{G}_i; \widehat{\Theta}^{(t)}\right) \tag{2.27}$$

$$\propto \tau_i^{(t)}(c, d; \widehat{p}^{(t)}, \widehat{q}^{(t)}) f\left(y_i; \widehat{\theta}_c^{(t)}, d\right). \tag{2.28}$$

In the M-step, we use these responsibilities to maximize the expectation of the complete

likelihood 2.26 under these posterior probabilities

$$\widehat{\Theta}^{(t+1)} = \underset{\Theta}{\text{argmax}}\ E_{C,\mathbf{1}'G|y,t,\widetilde{G},\widehat{\Theta}^{(t)}}\left[\sum_{i=1}^{N}\log\left[\tau_i(c,d;p,q)f\left(y_i;\theta_c,d\right)\right]\right] \tag{2.29}$$

$$= \underset{\Theta}{\text{argmax}}\ \sum_{i=1}^{N}\sum_{c}\sum_{d=0}^{N-1}\widehat{\gamma}_{icd}^{(t+1)}\log\left[\tau_i(c,d;p,q)f\left(y_i;\theta_c,d\right)\right]. \tag{2.30}$$

For example, under the linear model 2.9, we can compute closed form updates for the regression parameters:

$$\widehat{\beta}_c^{(t+1)} = \frac{\overline{dy}_c - \overline{d}_c\overline{y}_c}{\overline{d^2}_c - \overline{d}_c^2} \tag{2.31}$$

$$\widehat{\alpha}_c^{(t+1)} = \overline{y}_c - \widehat{\beta}_c^{(t+1)}\overline{d}_c \tag{2.32}$$

$$\widehat{\sigma}^{2(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c,d}\left[\widehat{\gamma}_{icd}^{(t+1)}\left(y - \widehat{\alpha}_c^{(t+1)} - \widehat{\beta}_c^{(t+1)}d\right)^2\right] \tag{2.33}$$

where $\overline{d}_c = \frac{\sum_{i=1}^{N}\widehat{\gamma}_{icd}^{(t+1)}d}{\sum_{i=1}^{N}\widehat{\gamma}_{icd}^{(t+1)}}$, $\overline{y}_c = \frac{\sum_{i=1}^{N}\widehat{\gamma}_{icd}^{(t+1)}y_i}{\sum_{i=1}^{N}\widehat{\gamma}_{icd}^{(t+1)}}$, $\overline{d^2}_c = \frac{\sum_{i=1}^{N}\widehat{\gamma}_{icd}^{(t+1)}d^2}{\sum_{i=1}^{N}\widehat{\gamma}_{icd}^{(t+1)}}$, and $\overline{dy}_c = \frac{\sum_{i=1}^{N}\widehat{\gamma}_{icd}^{(t+1)}dy_i}{\sum_{i=1}^{N}\widehat{\gamma}_{icd}^{(t+1)}}$. Note the similarly of our linear model estimators to those obtained from weighted least squares. Updates for the mismeasurement parameters $p$ and $q$ cannot be computed in closed form but can be solved for using a general optimizer such as `optim` in R. Estimates of the mean outcomes under each exposure condition (2.24) are functions of the model parameters and can be calculated accordingly.

Given the likelihood (2.26) is bounded and satisfies mild smoothness conditions, as well as sufficiently many runs of the EM algorithm, we should be able to find the global optima and obtain the MLE $\widehat{\Theta}_{EM} = \widehat{\Theta}_{MLE}$ (McLachlan and Basford, 1988). We can consider the consistency of $\widehat{\Theta}_{MLE}$ under the scenario we had access to comparable experiments on many networks, and that the outcomes for each experiment arise from the mixture model (2.21) with the same set of parameters $\Theta^*$. The major condition from Wald (1949) needed to ensure consistency of $\widehat{\Theta}_{MLE}$ is the identifiability of the mixture model on a non-zero probability set of the subjects in these experiments. In the case of our linear model for the potential

outcomes (2.9), consistency requires regular variation in the observed in-degrees and exposure conditions. Building off of Proposition 1, a sufficient condition for the consistency of $\widehat{\Theta}_{MLE}$ would be to observe infinitely many subjects with at least two distinct observed in-degrees $\tilde{d}$ under each observed exposure condition.

While the EM algorithm does not provide standard errors and confidence intervals for our estimate $\widehat{\Theta}_{EM}$ and functions thereof (such as the mean outcome estimates), bootstrap methods have been used in the context of other mixture models to approximate these quantities (Basford et al., 1997). In particular, we consider the parametric bootstrap, which consists of the following steps:

1. Generate samples $\{y^{(1)}, y^{(2)}, .., y^{(m)}\}$ from the fitted model given by (2.21) with parameters $\widehat{\Theta}_{EM}$, holding the treatment assignment vector $t$ and observed network $\widetilde{G}$ fixed.

2. Estimate $m$ sets of parameters $\{\widehat{\Theta}^{(1)}_{EM}, \widehat{\Theta}^{(2)}_{EM}, .., \widehat{\Theta}^{(m)}_{EM}\}$ using the EM algorithm on the generated samples $\{y^{(1)}, y^{(2)}, .., y^{(m)}\}$.

3. Calculate Monte-Carlo estimates of the standard errors and/or confidence intervals for $\widehat{\Theta}_{EM}$ using the parameters estimates $\{\widehat{\Theta}^{(1)}_{EM}, \widehat{\Theta}^{(2)}_{EM}, .., \widehat{\Theta}^{(m)}_{EM}\}$.

### *2.4.4 Simulation Study*

In this section, we apply our mixture model to simulated experiments run over the households of 75 Indian villages, collected by and described in detail in (Banerjee et al., 2013). Within each village, we consider an experiment with two treatment levels ("treatment" and "control") and treatment interference on the outcome of interest through the household network of borrowing and lending money. [14]

[14] Data are available from `https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/21538`.

The 75 villages range in size from 77 to 356 households, with an average of slightly under 200. We take the true spillover network in each village to be the reported network of borrowing and lending, with a link between two households if there is any monetary borrowing or lending between the two. While these networks may themselves be mismeasurements of the actual borrowing/lending networks for each village, they are nonetheless a useful laboratory in which to explore the performance of our proposed estimator. In particular, they exhibit realistic properties one may expect in networks, such as small world phenomenon, significant clustering, and substantial variation in degrees. When excluding households with zero degree, the average number of households in a village is about 170.

We run 10 simulated experiments on each village, each of which consists of the following steps:

1. Independently assign each household to treatment with probability 0.25.

2. Calculate the exposure condition $C_i$ for each household $i$ using the true network $G$.

3. Generate outcome of interest $y_i$ for each household according to (2.9) with parameters $\alpha = (0, 0.25, 0.5, 1)$, $\beta = (0.05, 0.1, 0.05, 0.1)$, and $\sigma^2 = 0.25$.

4. Observe a mismeasured network $\widetilde{G}$, where each edge in the true network is observed independently with probability $1 - p$, while false edges in the network are observed independently with probability $qd$, where $d$ is the density of true graph $G$[15]. Simulations are repeated for every pair of mismeasurement probabilities $(p, q)$ with $p \in \{0, \frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}\}$ and $q \in \{0, \frac{1}{8}, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}\}$.

Across the 10 simulations and 75 villages, we average 45 households with no exposure to treatment (under the true network), 82 households with indirect exposure to treatment, 15 households with direct exposure to treatment, and 28 households with full exposure to

[15]This specification for the mismeasurement parameter governing the observance of false edges is easier to compare across networks with varying densities. $q = 0.5$ implies observing a number of false edges equal to about half the actual number of true edges.

treatment. The challenge is to derive estimates of the mean outcome under each exposure condition (2.5) despite observing a mismeasured network $\widetilde{G}$ and thus mismeasured exposure conditions. Some of our exposure conditions can be quite challenging; for example, when $p = q = 0.5$, on average half of the true edges are not observed, but instead are "replaced" with roughly the same number of false edges.

Estimation proceeds as described in Section 2.4.3. We choose the parameters of the beta-binomial distribution over the true degrees (recall the discussion following equations 2.16 and 2.17) by taking $\mu$ to be the density of the true network and choosing dispersion parameter $\rho$ such that the second moment of the beta-binomial distributions matches that of the the observed degree distribution. This simulation scenario is consistent with a setting where we have *a priori* expectations on the density of the true network of spillovers. In Figure 2.5 located in Appendix 2.7, we present results where $\mu$ and $\rho$ are chosen solely by matching the first two moments of the observed degree distribution. Our results are slightly worse under the purely empirical specification but still represent a substantial improvement over the Horvitz-Thompson approach that does not account for any mismeasurement.

In Figures 2.1 and 2.2, we present heat maps to compare the estimates obtained by the EM algorithm to the Horvitz-Thompson estimates that do not take into account missingness in the network, at varying levels of unreported true links ($p$) and falsely observed links ($q$). Positive deviations from the true mean outcomes are denoted in red while negative deviations are denoted in blue, with the intensity of the colors corresponding to the size of the deviation.

We first comment on bias in the Horvitz-Thompson estimates, corresponding to the analytical bias formula described in section 2.3. When we fix $q$ and increase the proportion of unreported true edges $p$, more links to treated subjects are dropped, leading to a larger proportion of subjects with indirect treatment to be falsely classified as not indirectly treated. The bias in the Horvitz-Thompson estimators for no exposure and direct exposure increase in magnitude, as these estimators trend towards the mean outcomes under the indirect and full exposure conditions respectively.

The bias in the HT estimators for indirect exposure and full exposure also increase in

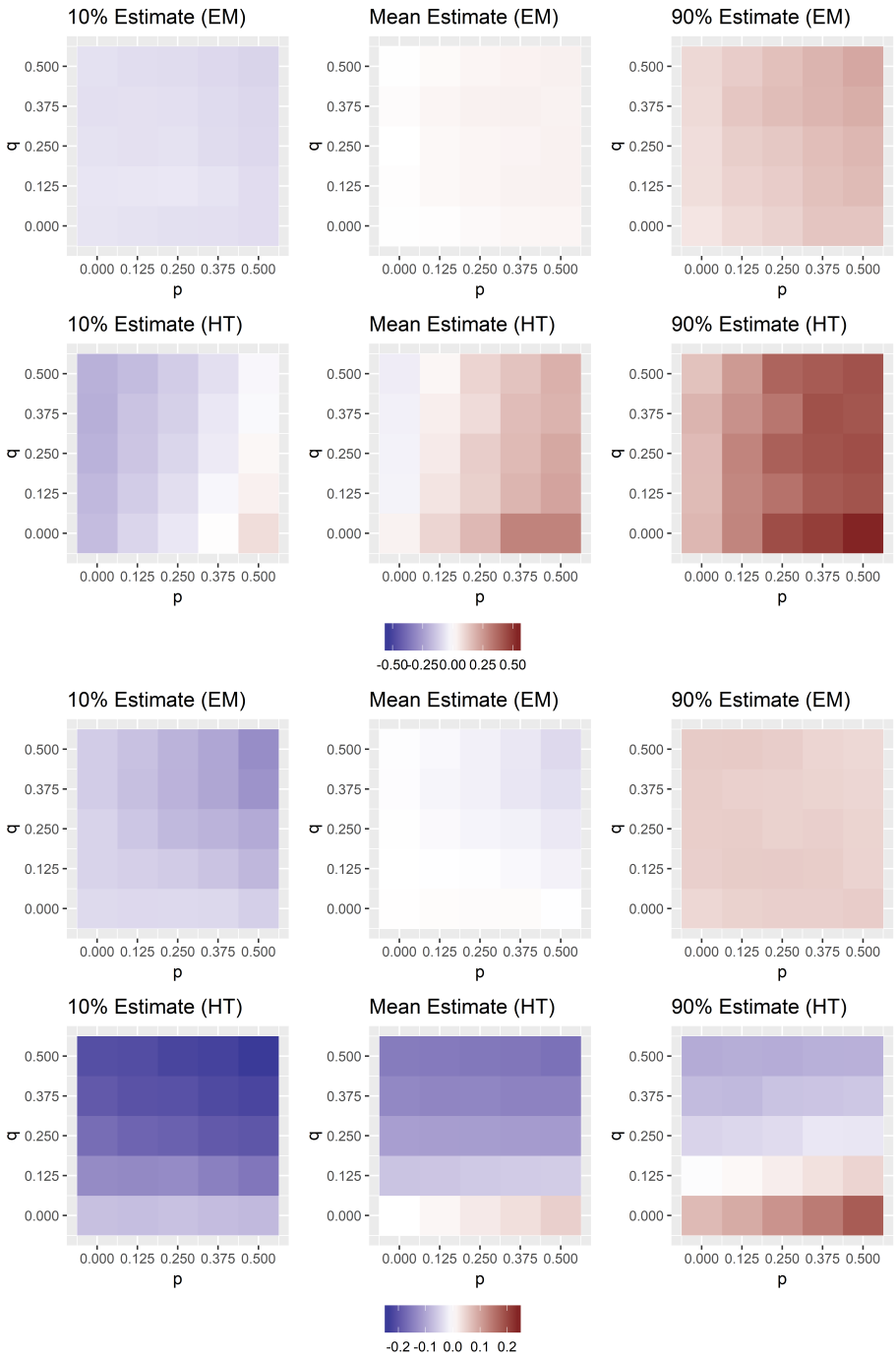


Figure 2.1: Estimates of the mean outcome of the no exposure (top) and indirect exposure (bottom) conditions from their true values under varying mismeasurement levels ($p$,$q$) for the network. Estimates obtained from our model using the EM algorithm are compared against estimates from the Horvitz-Thompson estimators assuming no mismeasurement in the network. The 0.1 and 0.9 quantiles are provided for both methods to give a sense of the variability in these estimates. Note the color gradient scales are different for the two exposure conditions.

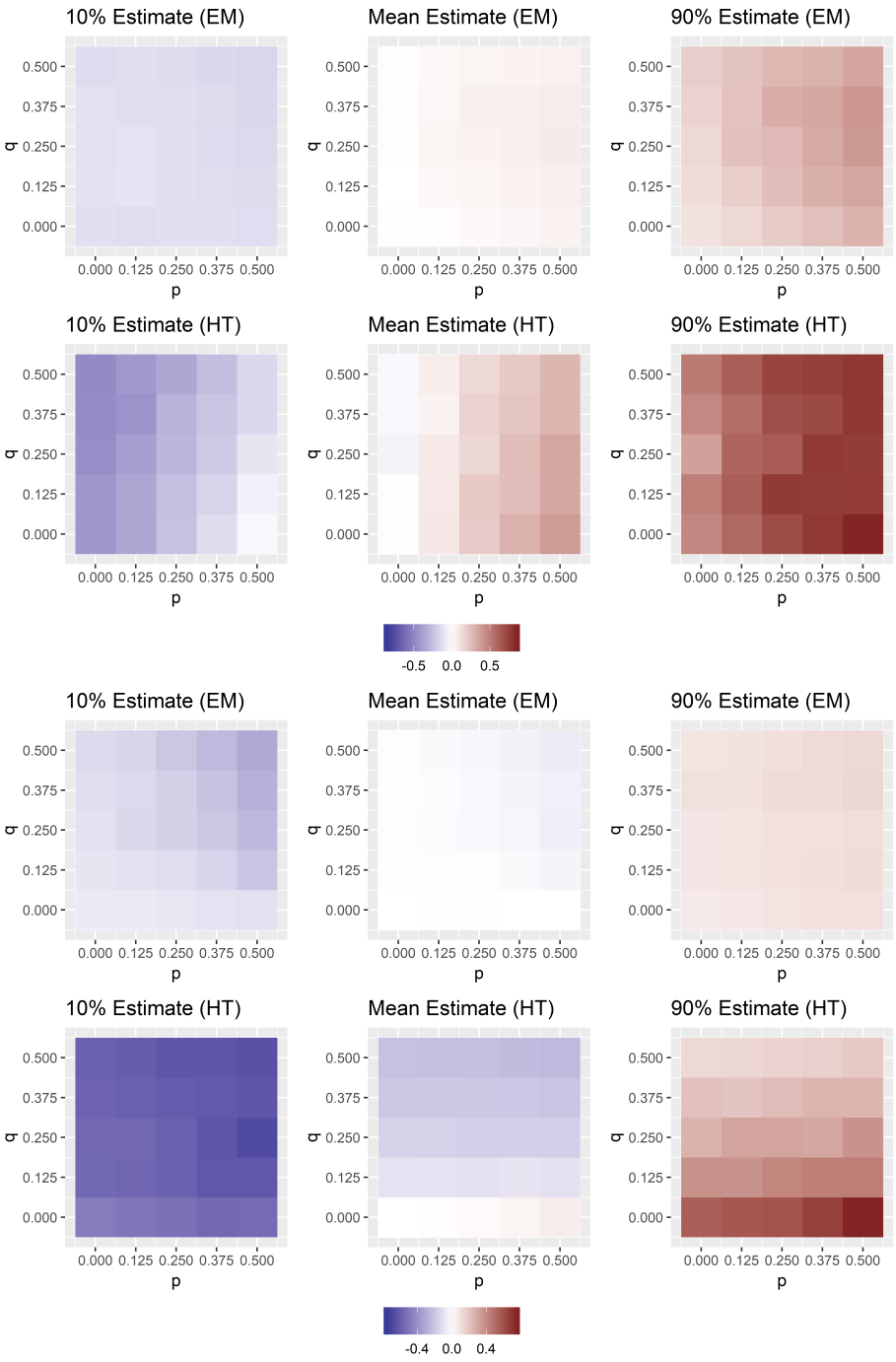


Figure 2.2: Estimates of the deviation of the mean outcome of the direct exposure (top) and full exposure (bottom) conditions from their true values under varying mismeasurement levels ($p$,$q$) for the network. Estimates obtained from our model using the EM algorithm are compared against estimates from the Horvitz-Thompson estimators assuming no mismeasurement in the network. The 0.1 and 0.9 quantiles are provided for both methods to give a sense of the variability in these estimates. Note the color gradient scales are different for the two exposure conditions.

magnitude, but to a lesser degree. This is because true links are dropped independently regardless of the treatment status of the subjects involved, the reduced number of subjects observed to be indirectly treated is accounted for by a decrease in the probabilities in the denominator of Equation 2.11. Rather, the bias for these estimators arises solely from low-degree subjects being removed from the estimation procedure (recall that the outcome model assumes that higher degree subjects tend to have higher outcomes). This effect is more pronounced at low $q$, when fewer (if any) false edges are being added and we are more likely to observe zero-degree subjects.

Similar patterns of behavior emerge when we fix $p$ and vary $q$. As $q$ increases, more subjects with no true connections to treated individuals are falsely observed to be indirectly treated. As a result, the HT estimators for indirect exposure and full exposure bias further downwards, towards the mean outcomes for no exposure and direct exposure respectively. Additionally, individuals with zero-degree are more likely to be included in the HT estimation procedure, biasing all four estimates downwards. Note this effect is not attenuated at higher $p$, since zero-degree individuals have no true links to drop to begin with.

By contrast, the estimates of mean outcome given by the EM algorithm are quite reasonable across the varying levels of mismeasurement $p$ and $q$ considered, and represent a substantial improvement over the Horvitz-Thompson estimates. Differences in performance across the various mismeasurement levels are much more muted, at least across the different levels of mismeasurement considered in our simulations.

To examine these results in more detail, we focus on the two cases presented in Figures 2.3 and 2.4, where we fix $p = 0.5$ and vary $q$ and fix $q = 0.5$ and vary $p$ respectively. Estimates from our method are presented in cyan while estimates from the Horvitz-Thompson are presented in red. In general, estimates from our method exhibit considerably less bias and are simultaneously have less variance. We find that our method has slightly higher levels of bias and variance for higher levels of mismeasurement, which is consistent with the idea that for higher $p$ and $q$ there is larger uncertainty over the true network (2.20) and thus our results are more dependent on the assumed beta-binomial model over the true degree

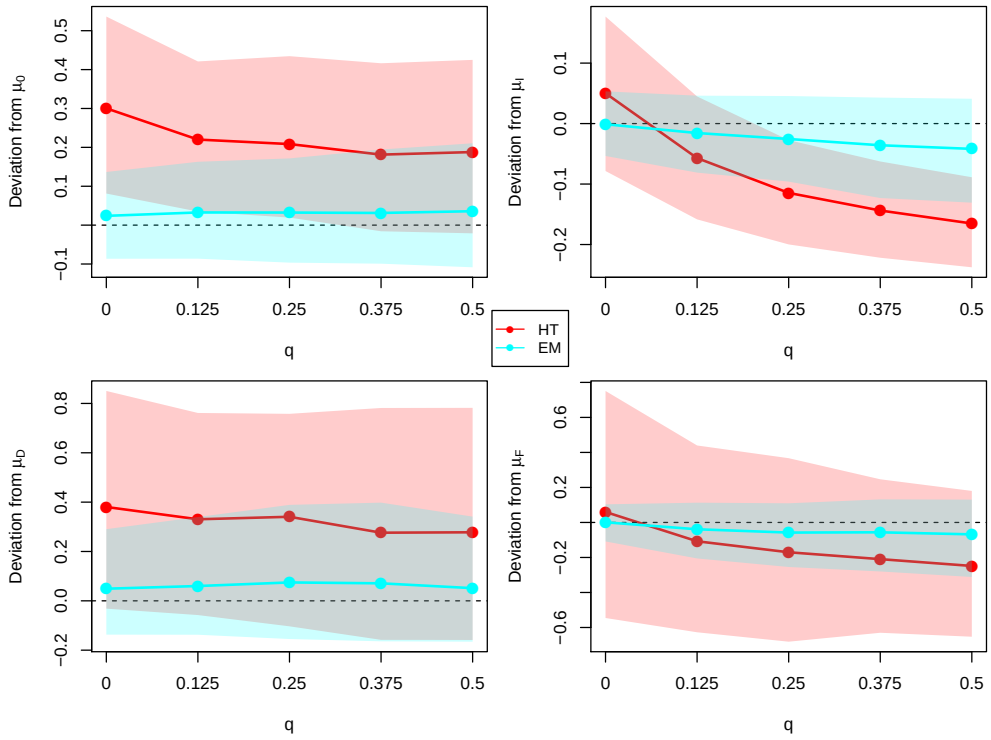


Figure 2.3: Estimates of the deviation of the mean outcome of each exposure conditions (top left: no exposure, top right: indirect exposure, bottom left: direct exposure, bottom right: full exposure) from their true values under $p = 0.5$ and varying $q$ from 0 to 0.5. Estimates obtained from our model using the EM algorithm are compared against estimates from the Horvitz-Thompson estimators assuming no mismeasurement in the network. The 0.1 and 0.9 quantiles are shaded for both methods to give a sense of the variability in these estimates.

distribution. Even if $\mu$ and $\rho$ are chosen to match the true degree distribution, the beta-binomial model still represents a (higher-order) deviation from the true degree distribution for real-life networks. The direction of the bias, upwards for $\widehat{\mu}_0$ and $\widehat{\mu}_D$ towards $\mu_I$ and $\mu_F$ respectively and the reverse for $\widehat{\mu}_I$ and $\widehat{\mu}_F$, are a product of imperfectly learning the latent exposure conditions, especially in these higher uncertainty settings.

## *2.5 Application to a study of diffusion of insurance information between farmers*

Cai et al. (2015) study the adoption decisions of rice farmers in rural China in regards to weather insurance, which typically exhibit low take-up rates even in the presence of heavy

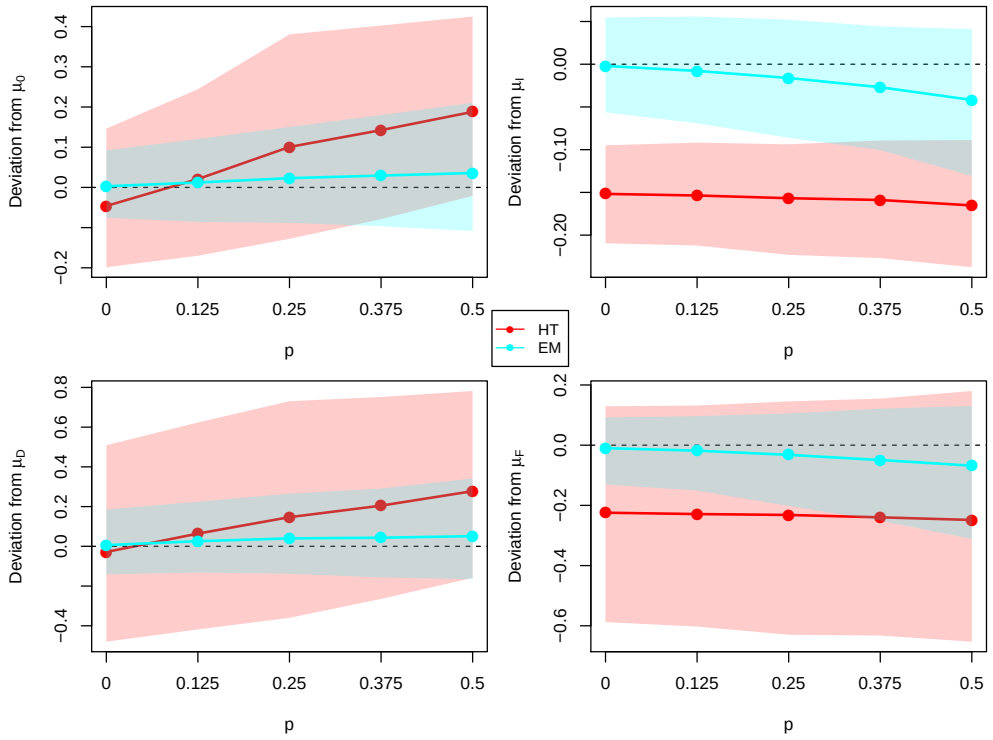


Figure 2.4: Estimates of the deviation of the mean outcome of each exposure conditions (top left: no exposure, top right: indirect exposure, bottom left: direct exposure, bottom right: full exposure) from their true values under $q = 0.5$ and varying $p$ from 0 to 0.5. Estimates obtained from our model using the EM algorithm are compared against estimates from the Horvitz-Thompson estimators assuming no mismeasurement in the network. The 0.1 and 0.9 quantiles are shaded for both methods to give a sense of the variability in these estimates.

government subsidies. The authors were interested in how difficulties communicating the benefits of the product may be modulated by the dissemination of information via a farmer's peers. In conjunction with the introduction of a new weather insurance product, researchers randomized about 5000 households across 185 rural villages into information sessions about the new insurance product. Sessions were held in two rounds three days apart, and could either be "simple" sessions just describing the product or longer "intensive" sessions which also emphasized the expected benefits from insurance. Drop out was not a major issue in this experiment, with an overall session attendance rate of about 90%.

One specific question the authors were interested in was how insurance take-up and

knowledge for households assigned to second round sessions were affected by whether or not they had friends assigned to first round intensive sessions. Before the experiment, each household was asked to list five friends whom they most frequently discussed production or financial issues with. About 96% list 5 friends included in the study, with the rest listing between one to four friends. In general, prompting respondents to list five friends can censor the number of connections for individuals with high in-degree, as well as cause the reported network to contain some weaker connections that would otherwise be unreported. However, this concern may be relatively mild in this case, as the authors conducted a pilot study in two villages where the number of friends was uncensored and found 96% of farmers reported either four or five connections. Most of the paper's results use this reported network, which, borrowing their language, we will term "general network measure."

Since insurance take-up is a binary outcome measure, which our method cannot handle [16] due to the unidentifiability of mixtures of Bernoulli variables, we focus on the network effect on insurance knowledge. To measure insurance knowledge, each household completed a five question test after the experiment and were scored from 0-5. The authors found that, for households assigned to second round sessions, having a friend who was assigned a first round intensive session had the same (statistically significant) benefit for insurance knowledge as personally being assigned an intensive session in the second round. Furthermore, households who had a friend who assigned to a first round intensive session derived no additional benefit in terms of insurance knowledge from being personally assigned to an intensive session.

First, we seek to replicate their finding with the following simple generative model:

$$\text{score} \sim \text{Bin}(5, \text{expit}(\alpha_c + \beta_c d)) \tag{2.34}$$

We model each household's score on the insurance test as arising from a binomial with five independent questions and a probability of getting a question right depending on the

[16] In Section 2.6, we provide a brief discussion of how our methodology might be modified to handle binary outcomes and other outcome distributions which are generically unidentifiable.

household's treatment exposure condition as well as their degree in the network. To produce comparable estimates to the linear specification (2) presented in Table 5 of their paper, we estimate the mean outcome under each exposure condition and calculate various contrasts using these means. The effect of personally being invited to a intensive session can be calculated as $\widehat{\mu}_D - \widehat{\mu}_0$, the effect of having a friend invited to a first round intensive session (denoted in our table as "Network Intensive") is calculated as $\widehat{\mu}_I - \widehat{\mu}_0$, and the interaction of these effects is given by $\widehat{\mu}_0 + \widehat{\mu}_F - \widehat{\mu}_I - \widehat{\mu}_D$. We present the results under the assumption the general network measure accurately characterizes the network of spillovers in Table 2.1, and compare those results to those obtained when allowing for potential mismeasurement in the reported network[17]. The two methods yield very similar estimates, as our method finds the outcomes consistent with a relatively low amount of mismeasurement in the observed network.

Next, we consider replicating the results using alternative network specifications. In particular, we consider two other network measures mentioned in Cai et al. (2015). The authors define a "strong" network measure where non-reciprocal connections are dropped, as well as a "weak" network measure which adds second-order connections ("friends of friends") to the general network measure. Both specifications differ quite drastically from the general network measure and are indicative of low rates of reciprocity and transitivity in the reported network; farmers average a single connection under the strong network measure (with a mode of 0) and 16 connections under the weak network measure. We repeat the estimation of direct and indirect treatment effects using these alternative network measures both when 1) assuming the network measure correctly specifies the spillover pathways and 2) allowing for mismeasurement in the network and report the results in Table 2.1. In contrast to the results assuming the network is correctly specified, our method gives pretty similar results under

[17]The beta-binomial distribution over the true degrees is initialized with the same mean as the observed network (reflecting the results from the pilot study) and overdispersion parameter 0.0005. This overdispersion parameter was chosen based on examining the variation of the degrees in the Indian village data used in the simulations. Note the observed degree-distribution in farmers' network exhibits considerably less variation than even a binomial distribution, and thus is not particularly informative for choosing our prior distribution.

| Network measure | Intensive Session | Network Intensive | Interaction |
|---|---|---|---|
| **No mismeasurement** | | | |
| General measure | 0.205 (0.016) | 0.198 (0.016) | -0.241 (0.023) |
| Strong measure | 0.100 (0.012) | 0.120 (0.036) | -0.188 (0.054) |
| Weak measure | 0.157 (0.061) | 0.072 (0.044) | -0.095 (0.063) |
| **EM method** | | | |
| General measure | 0.177 (0.025) | 0.229 (0.028) | -0.224 (0.040) |
| Strong measure | 0.299 (0.037) | 0.279 (0.040) | -0.577 (0.052) |
| Weak measure | 0.160 (0.032) | 0.259 (0.032) | -0.171 (0.048) |
| **EM + covariates** | | | |
| General measure | 0.158 (0.035) | 0.295 (0.039) | -0.322 (0.055) |
| Strong measure | 0.177 (0.081) | 0.305 (0.070) | -0.479 (0.116) |
| Weak measure | 0.175 (0.095) | 0.293 (0.067) | -0.155 (0.117) |

Table 2.1: Differences in insurance knowledge for farmers assigned to second round sessions based on (1) whether they attended an intensive session, (2) whether they had a friend attend a first round intensive session, and (3) the interaction of these two terms. We compare our method to results if we assume there is no mismeasurement in the network for three network measures.

the general and weak measures, despite large differences in the networks being used. We perform quite differently for the strong measure, although perhaps this should be expected given the sparsity of network information in this case.

Finally, we consider an extension of our model (2.20) that allows for different levels of mismeasurement in the connections between farmers depending on whether or not they reside in the same village. About 99.4% of reported connections are between farmers in the same village[18], while the remaining 0.6%, so separately modeling the true degree within-village and out-of-village may lead to more accurate results. We introduce distinct parameters for in-village degree ($\mu_{in}$ and $\rho_{in}$) and out-of-village degree ($\mu_{out}$ and $\rho_{out}$), along with respective mismeasurement parameters $p_{in}$, $q_{in}$, $p_{out}$, and $q_{out}$. Note that there is a potential variance trade-off when introducing additional parameters to our model, so sample size concerns must

[18]There is substantial variation in the size of each village, which is entirely not reflected under the network measures considered

also be considered. Estimation proceeds as described in Section 2.4.3, with the additional complication that the general purpose optimizer must maximize over all four mismeasurement parameters at once. Estimates from this extension are largely similar to those obtained from our method ignoring the difference between in-village and out-of-village ties, perhaps due to the lack of out-of-village ties. However, the corresponding standard errors are substantially larger.

## *2.6 Discussion*

Experimental inference on social networks present distinct challenges; not only are subjects' outcomes affected by the treatment assignments of other subjects, but this treatment interference is often of direct interest. Existing methodology for estimating treatment effects in this setting requires a precise measurement of the network of interest, which can be a difficult assumption given the many decisions inherent in the data gathering process as well as imposing a large financial burden. In this chapter, we present a class of mixture models that can accurately estimate treatment effects when the network of interest is not accurately measured, assuming the noise in the network is (conditionally) random and relying on additional assumptions about the parametric form for the treatment exposure conditions and the density of the true, latent network.

In many economic settings such as Cai et al. (2015), the primary behavior of interest may concern a binary outcome, such as the adoption decision of some product or technology. Due to the generic unidentifiability of the Bernoulli distribution, our mixture modeling framework is inherently unable to provide model parameter estimates in these settings. One set of additional assumptions in order to identify the mixture model given by (2.21) would be to fully specify the network corruption mechanism, i.e. choose the corruption probabilities $p$ and $q$. This set of assumptions, which may be relatively tenable if the corruption probabilities can be roughly ascertained via other sources of data, would result in a framework in which the uncertainty over the underlying network of interest is first modeled using the

provided corruption mechanism[19], and then treatment effects are estimated incorporating this uncertainty. In contrast, in the setting of identifiable outcome distributions, our method aims to jointly model the uncertainty in the latent network and in the treatment effects.

## 2.7 Additional Figures

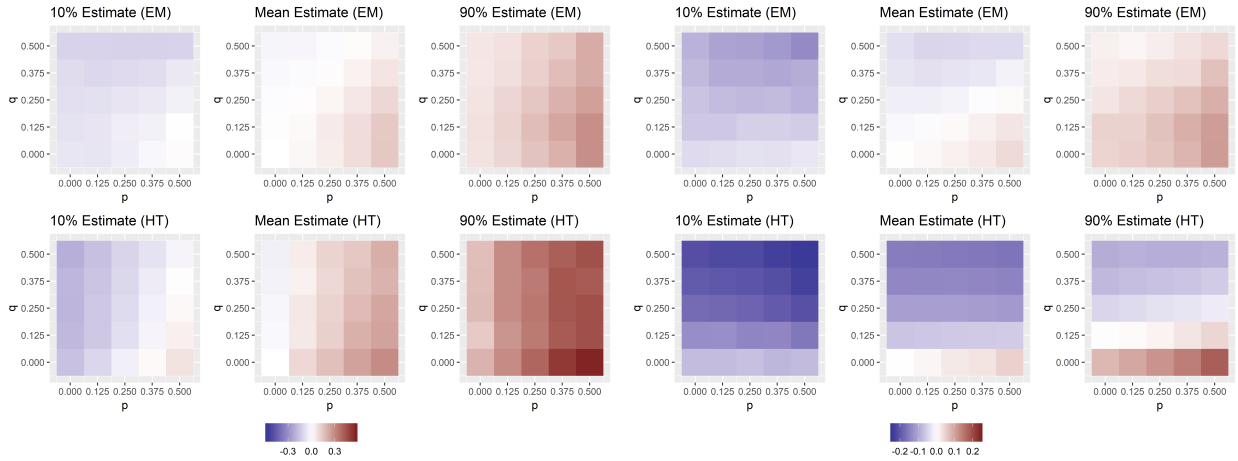


Figure 2.5: Estimates of the mean outcome of the no exposure (left) and indirect exposure (right) conditions from their true values under varying mismeasurement levels ($p$,$q$) for the network. Estimates obtained from our model using the EM algorithm are compared against estimates from the Horvitz-Thompson estimators assuming no mismeasurement in the network. The 0.1 and 0.9 quantiles are provided for both methods to give a sense of the variability in these estimates. Note the color gradient scales are different for the two exposure conditions.

## 2.8 Proof of Identification

It suffices to show identifiability of $\{\theta_{00}, \theta_{01}\} = \{\theta'_{00}, \theta'_{01}\}$, since we assume direct treatment status can always be accurately ascertained. The exposure conditions $\{c_{00}, c_{01}\}$ are only mismeasured with one another, as are $\{c_{10}, c_{11}\}$.

Let us begin with the most general case, when both $\{p, q\} \in (0, 1)$. In this situation, the probabilities $\tau_i$ are positive over all feasible true exposure conditions and degrees, regardless of

[19]This process is equivalent to sampling over the true latent network and has been applied in the context of network sampling (for example, see Griffin et al. (2018))

the pair of observed degrees ($\tilde{d}_t$,$\tilde{d}_{nt}$). The only restriction on the support of these probabilities are that, under no indirect treatment, degree cannot be larger than $N-1-1't+t_i \equiv N_{nt,i}$ (otherwise there would have to exist a connection to a treated subject), and degree must be at least one for an individual to be indirectly treated. Mathematically, $\tau_i(c_{00}, d; p, q) > 0$ for any $d$ satisfying $d \leq N_{nt,i}$ and $\tau_i(c_{01}, d; p, q) > 0$ for any $d \geq 1$.

At the other extreme, when there is no mismeasurement ($p = 0$ and $q = 0$), then the true exposure condition and degree match their observed counterparts. Mathematically, $\tau_i(c, d; p, q) > 0$ only for $d = \tilde{d}_t + \tilde{d}_{nt}$ and either $c = c_{00}$ if $\tilde{d}_t = 0$ or $c = c_{01}$ if $\tilde{d}_t > 0$. When exactly one kind of mismeasurement exists, the support of $\tau_i$ is limited, but to a lesser extent that when neither types of mismeasurement exist. When $p > 0$ but $q = 0$, true edges can be dropped but all observed edges also exist in the true network. Namely, any observed connection to a treated subject must exist in the true graph. For subjects with at least one of these connections $\tilde{d}_t > 0$, the support of $\tau_i$ is limited to $c = c_{01}$ and $d \geq \tilde{d}_t + \tilde{d}_{nt}$. If instead we have $\tilde{d}_t = 0$, $\tau_i$ is positive for $\tilde{d}_{nt} \leq d \leq N_{nt,i}$ when $c = c_{00}$ and $d \geq \tilde{d}_{nt} + 1$ when $c = c_{01}$. Lastly, when $q > 0$ and $p = 0$, the observed connections is a superset of the links in the true graph. Thus, when we observe no connections to treated subjects $\tilde{d}_t = 0$, $\tau_i$ is only positive for $c = c_{00}$ and $d \leq \tilde{d}_{nt}$. When such an connection is observed, $\tau_i$ is positive for $c = c_{00}$ and $d \leq \tilde{d}_t + \tilde{d}_{nt} - 1$ or $c = c_{01}$ and $1 \leq d \leq \tilde{d}_t + \tilde{d}_{nt}$.

Case 1: $p > 0$, $q > 0$, and $\beta_{10} \neq 0$

For any pair of $\left(\tilde{d}_t, \tilde{d}_{nt}\right)$, the LHS is a mixture of normal distributions that includes $N-1$ distinct components with means $\alpha_{01} + \beta_{01} d$ and variance $\sigma^2$ for any $d$ from $\{1, ..., N-1\}$. There are at most $N_{nt,i} + 1 < N - 1$ other mixture components corresponding to the $c_{00}$ terms. Following the generic identifiability of finite normal mixtures, the same component normals must exist on the RHS, with the same weights. For there to be at least $N-1$ distinct components on the RHS for both $\tilde{d}_t = 0$ and $\tilde{d}_t > 0$, we must have $p' > 0$ and $q' > 0$. On the LHS, we have $N - 1$ components which are evenly spaced $|\beta_{01}|$ apart, while on the RHS we have $N - 1$ components evenly spaced $|\beta'_{01}|$ apart. Since there are fewer than $N - 1$ other

components on either side, these $N-1$ components must match, with $|\beta_{01}| = |\beta'_{01}|$. This leads to two possibilities: we must have either $\alpha'_{01} = \alpha_{01}$ and $\beta'_{01} = \beta_{01}$ or $\alpha'_{01} = \alpha_{01} + N\beta_{01}$ and $\beta'_{01} = -\beta_{01}$. The latter cannot occur due to would-be inconsistencies in the weights. For example, consider weights for the component with mean $\alpha_{01} + \beta_{01}$ under this scenario. On the LHS, the weight would correspond to the probability $\tau_i(c_{01}, 1; p, q)$, while on the RHS, the weight would correspond to the probability $\tau_i(c_{01}, N-1; p', q')$. The former quantity changes with $\tilde{d}_t$ if holding the total observed degree $\tilde{d}_t + \tilde{d}_{nt}$ fixed, since the observed treated degree would affect the probability of a true treated connection, but the latter does not since for very large true degree $d > N_{nt,i}$ we will always have a treated connection. Thus, we have $\alpha'_{01} = \alpha_{01}$ and $\beta'_{01} = \beta_{01}$.

We can then use our identification of the $c_{01}$ components to isolate the remaining, unexplained components, which must correspond to $c_{00}$. If $\beta_{00} \neq 0$, the LHS has $N_{nt,i} + 1$ remaining components, while if $\beta_{00} = 0$, the LHS has one component. The same holds for $\beta'_{00}$ and the RHS. Thus, when $\beta_{00} = 0$, $\beta'_{00} = 0$ and we must have $\alpha'_{00} = \alpha_{00}$. On the other hand, if $\beta_{00} \neq 0$, both sides consist of $N_{nt,i} + 1$ components, spaced $|\beta_{00}|$ and $|\beta'_{00}|$ apart respectively. We must have either $\alpha'_{00} = \alpha_{00}$ and $\beta'_{00} = \beta_{00}$ or $\alpha'_{00} = \alpha_{00} + N_{nt,i}\beta_{00}$ and $\beta'_{00} = -\beta_{00}$. Following similar logic as above for the $c_{01}$ components, we can use would-be inconsistencies in the weights to eliminate the second scenario. Namely, consider the weights for the $\alpha_{00}$ component, which is $\tau_i(c_{00}, 0; p, q)$ for the LHS and $\tau_i(c_{00}, N_{nt,i}; p', q')$ for the RHS. For fixed $\tilde{d}_t + \tilde{d}_{nt}$, $\tau_i(c_{00}, 0; p, q)$ is unaffected by varying $\tilde{d}_t$ as all observed connections regardless of treatment status must be falsely observed, while the treatment status of the observed connections will effect the probability of having a treated connection given $N_{nt,i}$ true connections.

Case 2: $p > 0$, $q > 0$, and $\beta_{10} = 0$

Next, let us consider the scenario when we have $\beta_{10} = 0$, but $\beta_{00} \neq 0$. For any pair of $\left(\tilde{d}_t, \tilde{d}_{nt}\right)$, the LHS is a normal mixture including $N_{nt,i} + 1$ or $N_{nt,i} + 2$ components with means $\alpha_{01}$ and $\alpha_{00} + \beta_{00}d$ and variance $\sigma^2$ for any $d$ from $\{0, ..., N_{nt,i}\}$. Since the number of

components does not change for any pair of observed degrees, we have $p' > 0$, $q' > 0$, and $\beta'_{10} = 0$. Following the same logic used in case 1 but reversing the order in which we consider the $c_{00}$ and $c_{01}$ components, we can show $\{\theta_{00}, \theta_{01}\} = \{\theta'_{00}, \theta'_{01}\}$.

The alternate scenario involves the case $\beta_{00} = 0$ and $\beta_{01} = 0$. Since we assume there is a non-zero indirect treatment effect ($\theta_{00} \neq \theta_{01}$), the LHS consists of a mixture of two normals with means $\alpha_{00}$ and $\alpha_{01}$. Following the generic identifiability of normal mixtures, the RHS must consist of two normals with the same means. In order for the RHS to have two mixture components regardless of observed degree $\left(\tilde{d}_t, \tilde{d}_{nt}\right)$, we must have $\beta'_{00} = 0$ and $\beta'_{01} = 0$ as well as non-zero mismeasurement in both $p'$ and $q'$. For $q' = 0$, $\tilde{d}_t > 0$ would yield just one mixture component, and similarly with $\tilde{d}_t = 0$ for $p' = 0$. Thus, either $\alpha'_{00} = \alpha_{00}$ and $\alpha'_{01} = \alpha_{01}$ or $\alpha'_{00} = \alpha_{01}$ and $\alpha'_{01} = \alpha_{00}$. If the latter is the case, the weight of the $\alpha_{00}$ component is the probability of no indirect treatment $\sum_d \tau_i(c_{00}, d; p, q)$ on the LHS and the probability of indirect treatment $\sum_d \tau_i(c_{01}, d; p', q')$ on the RHS. These weights must be the same for any pair of $\left(\tilde{d}_t, \tilde{d}_{nt}\right)$. However, when holding $\tilde{d}_t + \tilde{d}_{nt}$ fixed and increasing the number of observed connections to treated individuals $\tilde{d}_t$, the weight of the LHS decreases while the weight of the RHS increase. Thus, we must have $\alpha'_{00} = \alpha_{00}$ and $\alpha'_{01} = \alpha_{01}$.

Case 3: $p > 0$ and $q = 0$

First, consider an observation $i$ with at least one observed connection to a treated subject $\tilde{d}_t > 0$. The mixture on the LHS consists of components with means $\alpha_{01} + \beta_{01} d$ corresponding to $\tau_i(c_{01}, d; p, q)$ for any $d$ satisfying $d \geq \tilde{d}_t + \tilde{d}_{nt}$. If $\beta_{01} = 0$, the LHS will just be one component, while if $\beta_{01} \neq 0$, the LHS will have $N - (\tilde{d}_t + \tilde{d}_{nt})$ distinct components.

Suppose for now the latter is true. Then increasing total observed degree $\tilde{d}_t + \tilde{d}_{nt}$ decreases the number of components on the LHS. Changing total observed degree has no effect on the number of distinct components when $p, q > 0$ or $p = q = 0$, while the case $p = 0$ and $q > 0$ would imply an increase in the number of distinct components. Thus, to match the behavior on the RHS, we must have $p' > 0$ and $q' = 0$. For the components on both sides to have the same set of means, we must have either $\alpha'_{01} = \alpha_{01}$ and $\beta'_{01} = \beta_{01}$ or

$\alpha'_{01} = \alpha_{01} + (N - 1 + \tilde{d}_t + \tilde{d}_{nt})\beta_{01})$ and $\beta'_{01} = -\beta_{01}$. We can again invalidate the second case by examining would-be inconsistencies in the weights $\tau_i$, but in this case we can also simply note that the latter scenario cannot be simultaneously valid across multiple choices of $\tilde{d}_t + \tilde{d}_{nt}$. Having established $\alpha'_{01} = \alpha_{01}$ and $\beta'_{01} = \beta_{01}$, we can consider observations with $\tilde{d}_t = 0$ and isolate the remaining $\tau_i(c_{00}, d; p, q)$ components on the LHS, of which there would be either 1 (if $\beta_{00} = 0$) or $N_{nt,i} - \tilde{d}_{nt} + 1$ (if $\beta_{00} \neq 0$) components. Matching these components on the RHS across multiples values of $\tilde{d}_{nt}$ will avoid the potential case where $\beta'_{00} = -\beta_{00}$ and yield $\alpha'_{00} = \alpha_{00}$ and $\beta'_{00} = \beta_{00}$.

Let us now return to the case where $\beta_{01} = 0$. While we could still find $\beta'_{01} = 0$ and $\alpha'_{01} = \alpha_{01}$, examining the number of components when $\tilde{d}_t > 0$ is not sufficient to imply $p' > 0$ and $q' = 0$. However, we can attempt to ascertain whether or not this must be the scenario by examining observations with $\tilde{d}_t = 0$. If $\beta_{00} \neq 0$, there would be $N_{nt,i} - \tilde{d}_{nt} + 1$ distinct components on the LHS. A decreasing number of components for these observations as $\tilde{d}_{nt}$ increase is only consistent with $p' > 0$ and $q' = 0$. From here, we can use the equal spacing of these components as well as the structure imposed by the weights to show $\alpha'_{00} = \alpha_{00}$ and $\beta'_{00} = \beta_{00}$.

Lastly, when both $\beta_{00} = 0$ and $\beta_{01} = 0$, we observe one mixture component with mean $\alpha_{01}$ when $\tilde{d}_t > 0$ and two mixture components with means $\alpha_{00}$ and $\alpha_{01}$ when $\tilde{d}_t = 0$. Returning to the logic used in the counterpart scenario in case 2, the LHS can only be matched when $p' > 0$ and $q' = 0$. Then the RHS will have one mixture component when $\tilde{d}_t > 0$ and two components when $\tilde{d}_t == 0$, and we will have $\{\theta_{00}, \theta_{01}\} = \{\theta'_{00}, \theta'_{01}\}$.

Case 4: $p = 0$ and $q > 0$

This case follows identical logic as case 3 but switching the roles of the $c_{00}$ and $c_{01}$ components. Namely, observations with $\tilde{d}_t = 0$ will isolate the $c_{00}$ components, which in turn can be used to inform observations with $\tilde{d}_t > 0$ to match the $c_{01}$ components.

Case 5: $p = 0$ and $q = 0$

For any pair of $\left(\tilde{d}_t, \tilde{d}_{nt}\right)$, the LHS will consist of consist of a single normal distribution. If $p' > 0$ or $q' > 0$, this behavior could only arise if $\beta_{00} = \beta_{01} = 0$ and $\alpha_{00} = \alpha_{01}$. However, we require $\theta'_{00} \neq \theta'_{11}$, so we must have $p' = 0$ and $q' = 0$. Observations from two distinct values of $\tilde{d}_t + \tilde{d}_{nt}$ for each of $\tilde{d}_t = 0$ and $\tilde{d}_t > 0$ will uniquely identify the model parameters $\theta_{00}$ and $\theta_{01}$ respectively.

# Chapter 3

# INFERRING SOCIAL STRUCTURE FROM CONTINUOUS-TIME INTERACTION DATA

Joint work with Bailey K. Fosdick and Tyler H. McCormick.

## *3.1 Introduction*

Relational event data present an attractive alternative to survey data, allowing for researchers to collect data on populations not accessible through surveys and circumvent mismeasurement issues inherent to the survey process. However, techniques for analyzing these sources of data have lagged behind, and instead researchers attempt to mirror the form of traditional survey network data by aggregating interactions into binary or weighted adjacency matrices, whose elements represent the presence of any interaction or number of interactions between each pair of actors. Analyses using this network, for example estimating treatment effects in experiments as in the setting of Chapter 2, would proceed under the assumption that the aggregated activity network would accurately represent the corresponding relational network. However, we argue that unobserved meaningful social (network) relations are not easily ascertained from the observed noisy interaction counts.

Furthermore, an aggregation-based approach can only accommodate the temporal element of relational event data in discrete-time, by aggregating interactions in fixed time intervals. Inference on resulting sequence of adjacency matrices is typically performed using models that characterize the manner in which dyadic interactions change from one time interval to the next, often assuming a Markov process whereby the network at time $t$ only depends on its past states through the network at time $t-1$ (Snijders, 2005; Sewell and Chen, 2015). A key drawback of performing discrete-time data aggregation when modeling

continuous-time interaction data is that the time intervals are often chosen arbitrarily and conclusions can be greatly impacted by these choices (Sulo et al., 2010; Sekara et al., 2016). In addition, the temporal dynamics of the interactions are possibly lost in the aggregation process depending on the length of the time interval. Consider Figure 3.1, for example, where proximity data from a barn swallow study has been aggregated over each day (see Sections 3.1.1 and 3.5.1 for further details on this data). By aggregating to the level of the day

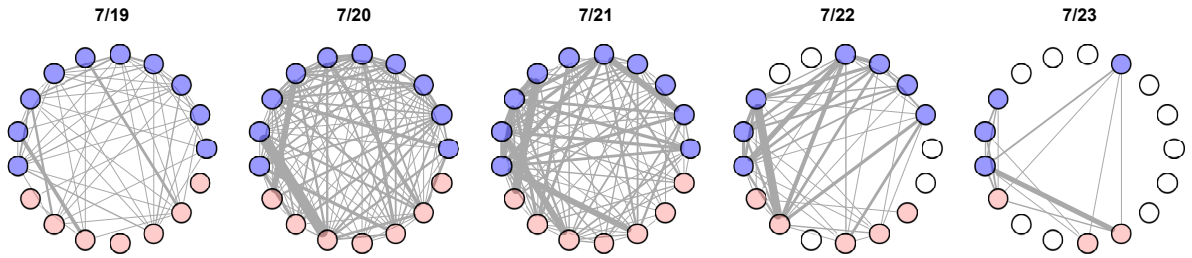


Figure 3.1: Daily snapshots of aggregated barn swallow interactions. Nodes are colored by sex (females red; males blue) and edges widths are proportional to the number of interactions between each pair. Birds deemed to be unmonitored on a given day (see Appendix 3.8.1) are shown in white.

we preclude the possibility of capturing intricate within-day social dynamics. While smaller time intervals are able to capture more of these dynamics, models with Markov structure may be inappropriate for modeling the resulting series of weighted networks. Conditional dependence among behavior across time would then be more likely to span multiple intervals. Returning to the example of call records, an assumption of Markovian dynamics for summaries of behavior at, for example, an hourly level would be unsuitable since associated individuals are unlikely to call each other on such a frequent basis.

In this chapter, we propose a continuous-time approach to modeling relational event data that explicitly separates interactions from underlying social relations. The model we propose possesses two distinctive properties. First, rather than viewing the data as direct observations of the network of interest, we assume observed interactions arise from a point processes with propensity influenced by the latent network structure. It is the dynamics of this *underlying* social network that we argue is typically most informative about population social structure

and of direct interest to researchers in inference settings such as Chapter 2. Second, we avoid decisions on temporal resolution by modeling both the observed interactions and dynamics of the latent network structure in continuous-time. The statistical challenge, therefore, is to infer the underlying network structure and its evolution through time. In this chapter we assume interactions and connection in the latent network are undirected, although our methods could readily be extended to handle directed interactions and networks.

Existing work in continuous-time modeling for relational data focuses on *excitatory* point processes (Simma and Jordan, 2010; Blundell et al., 2012; Perry and Wolfe, 2013). Hawkes processes (Hawkes, 1971) are one such process; in a Hawkes process, the occurrence of an event causes an immediate increase in intensity for that process and/or other Hawkes processes. These processes are favored for their ability to model reciprocity and transitivity in relational event data. Despite being used to model continuous-time data, existing point process models assume a static underlying network (Blundell et al., 2012; Linderman and Adams, 2014) or eschew modeling a network altogether, focusing on just modeling the interaction data (DuBois and Smyth, 2010; Simma and Jordan, 2010; Perry and Wolfe, 2013).

In this chapter, our primary focus is on inferring an underlying dynamic network, as opposed to inferring casual dependencies between the noisy interactions. In this sense, our work parallels that of (Linderman and Adams, 2014) and (Scharf et al., 2016), which aim to infer network relations from individual-level movement and behavior. We posit that the patterns in relational events are primarily governed by temporal and covariate-dependent factors and only secondarily by excitatory triggers. Thus, we model relational events as arising from inhomogeneous Poisson processes with intensities dependent on time, covariates, and an underlying dynamic latent network. In the interest of parsimony, if two individuals lack a relationship in the underlying network, we consider their interactions as the result of spurious behavior and thus use a relatively simple model. Model complexity is instead focused on interactions between pairs of individuals inferred to have a connection in the underlying network.

The remainder of the chapter is organized as follows. In Section 3.1.1, we briefly introduce

two applications for our relational event framework involving face-to-face interactions among college students and animal telemetry data. In Section 3.2, we describe the general framework in detail, and in Section 3.3 we propose a Bayesian inference procedure for models from this framework. We apply our framework to the two aforementioned datasets in Sections 3.4 and 3.5. Codes to replicate the analyses and figures presented in this chapter are available at `https://wesleytlee.github.io/relational-event-networks/`.

### *3.1.1 Proximity interactions: students and swallows*

Using our method, we explore network structure in two settings with data collected using proximity sensors: one involving interactions between college students and another consisting of contacts among barn swallows (*Hirundo rustica erythrogaster*).

First, we consider proximity information for 57 undergraduate students living in a dormitory at MIT during the 2008-2009 school year (Madan et al., 2012). Participants were given Bluetooth-enabled smartphones, and time stamps of Bluetooth (proximity) pings between nearby phones were recorded. In addition, students were surveyed five times throughout the school year and asked about assorted measures of their health and relationships with other subjects in the study. This dataset has been used to study the spread of infection (Dong et al., 2012) and the proliferation of health-related choices among these undergraduates (Madan et al., 2010). In both studies, researchers found physical contact information derived from the proximity data played a critical explanatory role beyond that of the self-reported relationships. These findings imply that information obtained from human behavior can inherently differ from that in reported relationships, and the relative importance and validity of each information type may vary across applications.

Next, we examine interactions within a population of Colorado barn swallows collected over three days during mating season (Levin et al., 2015, 2016). Barn swallows are small birds often found nesting in man-made structures. Due to their small size and speed in flight, recording proximity-based social interactions by visual observation is infeasible. Proximity loggers have emerged as promising alternative tools for collecting animal interaction data

in species of all sizes. The deployment of these tools presents new opportunities to gain understanding of animal social structure. We investigate a set of close encounters between barn swallows outfitted with proximity loggers.

### *3.2 A Continuous-Time Interaction Framework*

We propose a hierarchical modeling framework for continuous-time relational event data in which dyad interactions are modeled using an inhomogeneous Poisson process with intensity dependent on dyadic covariates and the state of the dyad in an underlying latent network. The latent network relation for each dyad is binary and modeled with a slow-changing continuous-time Markov chain. This construction allows our framework to express two properties of primary interest: 1) We avoid directly equating interaction frequency with relational strength in the network by detaching the process of modeling the relational events from the process of modeling the social network. Consequently, actors that are connected in the latent network can assume interaction patterns with varying frequencies. 2) Both the interactions and the network are modeled directly in continuous-time so that no choices about temporal resolution of data aggregation are required. Furthermore, we retain both the flexibility to capture fine temporal dynamics and the computational benefits of modeling the network evolution as Markovian.

For every pair of actors $(i, j)$, denote the vector of interaction event times as $\boldsymbol{t}^{ij} \equiv \left\{t_1^{ij}, ..., t_{N_{ij}}^{ij}\right\}$ and the time interval on which these events have the potential to occur by $\boldsymbol{T}^{ij} \equiv \left[T_1^{ij}, T_2^{ij}\right)$. Additionally, denote the set of dyadic, possibly time-varying, covariates as $x_{ij}(t)$. For the MIT data, covariates of potential interest include whether students reside on the same floor of the dormitory or are in the same year of school. For the swallow data, covariates include sex pairings and similarities between phenotypic traits.

We model the vector of interaction times $\boldsymbol{t}^{ij}$ for dyad $(i, j)$ using an inhomogeneous Poisson process with intensity $\lambda_{ij}(t)$, representing the instantaneous rate of an interaction occurring between actors $i$ and $j$ at time $t$. The log probability of events occurring at times

$\boldsymbol{t}^{ij}$ for dyad $(i, j)$ is then given by

$$\log p\Big(\boldsymbol{t}^{ij}\big|\lambda_{ij}(t)\Big) = \sum_{t\in\boldsymbol{t}^{ij}} \log \lambda_{ij}(t) - \int_{\boldsymbol{T}^{ij}} \lambda_{ij}(t)dt. \tag{3.1}$$

Conditional on the underlying intensity function, the presence/absence of interactions in disjoint time intervals are independent. Additionally, we assume interactions for different dyads arise independently conditional on their intensity functions. In contrast to Hawkes processes, dependence between dyad interactions is therefore entirely captured by the dyad intensities. Using $y_{ij}(t)$ as an indicator of whether a network connection exists between $i$ and $j$ at time $t$, we model each intensity $\lambda_{ij}(t)$ as dependent on the time $t$, dyad-specific covariates $x_{ij}(t)$, and the latent connection $y_{ij}(t)$.

Introducing time and covariate dependencies of the event data into the Poisson intensity function modulates the influence of these factors in the estimated network and allows the model to differentially weigh interactions as evidence of network structure. Interactions occurring at specific times and/or under certain circumstances may be much more probable in the presence of a network connection rather than in the absence thereof, and should be accounted for accordingly. We structure $\lambda_{ij}(t)$ as follows:

$$\lambda_{ij}(t) = \Bigg(1 + m\Big(t, x_{ij}(t)\Big)y_{ij}(t)\Bigg) \times \Bigg[w\Big(t, x_{ij}(t)\Big) + a\Big(t, x_{ij}(t)\Big)y_{ij}(t)\Bigg], \tag{3.2}$$

and denote the corresponding sets of model parameters for functions $\{m, w, a\}$ as $\{c_m, c_w, c_a\}$. Assuming no connection in the latent network, $y_{ij}(t) = 0$, the intensity function reduces to a baseline intensity $w(t, x_{ij}(t))$, dependent on time and dyad-specific covariates. The function $w$ corresponds to the rate of "spurious" interactions, which result when actors interact due to circumstance. For example, students in the MIT dormitory may be in close proximity to one another at night due solely to room assignments. We leave the exact form of $w$, as well as the other functions, to be specified on an case-by-case basis using domain expertise and descriptive summaries of the data.

Behavior for connected dyads can differ from the baseline in two ways: they may generally tend to interact more often, through the multiplicative factor $m$, and their interactions may have a different pattern, through an additive adjustment to the baseline $a$. When inferring the latent network, interactions in line with expected behavior for connected dyads but not expected behavior for unconnected dyads are suggestive of a latent connection. Thus, both evidence of generally higher activity levels than baseline and interactions that deviate from the baseline pattern according to our adjustment $a$ suggest an underlying relationship.

Poisson processes model interactions as instantaneous events in time. However, for many potential applications, including both the MIT and barn swallow proximity data, encounters can last for non-negligible durations. In these cases, we remove the time while the event is taking place from the time interval $\boldsymbol{T}^{ij}$ for each given pair, preventing the model from modeling the probability of another event while an interaction is taking place and assuring that the intensity is calibrated to the appropriate time window. Future work could extend our modeling framework to use the duration length as a potential source of information about future interactions and/or network relations.

The latent network $\boldsymbol{Y} \equiv \{y_{ij}(t) : i, j \in \{1, ..., n\},\, i \neq j,\, t \in \boldsymbol{T}^{ij}\}$ can be decomposed into dyad-specific paths $\{y_{ij}(t) : t \in \boldsymbol{T}^{ij}\}$ representing the state of a connection between dyad $(i, j)$ over time. These paths $\{y_{ij}(t)\}$ are modeled as independent stationary continuous-time Markov chains (CTMCs) taking values in $\{0, 1\}$. Each path $y_{ij}(t)$ is characterized by a sparsity parameter $s > 0$ and an transition parameter $q > 0$. The probability of transitioning from one network state to another in any interval of length $t$ is given by the entries of the following transition matrix:

$$P = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} (1-s) + se^{-qt} & s - se^{-qt} \\ (1-s) - (1-s)\, e^{-qt} & s + (1-s)\, e^{-qt} \end{bmatrix} \end{array}. \tag{3.3}$$

This Markov chain has stationary distribution equal to $(1 - s, s)$, meaning that, asymp-

totically, $y_{ij}(t) = 1$ proportion $s$ of the time. Thus, the parameter $s$ governs the overall propensity for a connection between any dyad, while $q$ governs the speed at which the network changes, i.e. relations form/dissolve. Modeling each binary relation as a CTMC with small $q$ values yields a slow-changing network.

In addition, we allow the CTMC parameters to depend on non-time-varying dyadic covariates, i.e. $s = s(x_{ij})$ and $q = q(x_{ij})$ with respective sets of parameters $\{c_s, c_q\}$. Allowing for the sparsity parameter to depend on dyadic covariates allows for the introduction of some network structure, such as block models. For example, one could cluster individuals in the inferred network into known communities (such as floors in the MIT data) by allowing the sparsity parameter to positively depend on whether or not individuals in each dyad shared the same community.

The hierarchical nature of the model allows us to separate the process of modeling a network from the process of modeling interactions. The interaction model acts as a low-pass filter on the data, distinguishing between spurious interactions and those suggestive of a meaningful relationship. Evidence of network relations between actors is more generally interpreted through deviance from the expected baseline level of activity rather than only through high raw interaction counts. The hidden CTMC under the interaction model smooths evidence across time so that relationships are thus characterized by consistent, prolonged deviations from baseline behavior.

### *3.3 Bayesian Inference*

We propose a Bayesian sampling procedure for estimating the parameters of the model $\boldsymbol{\theta} \equiv \{c_m, c_w, c_a, c_s, c_q\}$ and the latent network $\boldsymbol{Y} = \{y_{ij}(t) : i, j \in \{1, ..., n\},\, i \neq j,\, t \in \boldsymbol{T}^{ij}\}$. After appropriately eliciting priors, we are interested in the posterior distribution of the parameters and network conditional on the data $D \equiv \{\boldsymbol{t}^{ij} : i, j \in \{1, ..., n\}\}$. Although the posterior distribution is analytically intractable, we can approximate it to an arbitrary degree of accuracy by taking samples from the posterior with a Markov chain Monte Carlo (MCMC) algorithm. MCMC algorithms involve constructing a Markov chain with stationary

distribution equal to the posterior distribution of the parameters and network given the data.

Inference proceeds by first sampling the parameters and then estimating the network. The posterior distribution of the parameters $\boldsymbol{\theta}$ given the data $D$ is defined

$$p\left(\boldsymbol{\theta}|D\right) \propto \left[\prod_{(i,j)} p\left(\boldsymbol{t}^{ij}|\boldsymbol{\theta}\right)\right] \times p\left(\boldsymbol{\theta}\right). \tag{3.4}$$

After obtaining a sample of size $K$ from this posterior, the posterior probability that dyad $(i, j)$ is connected in the network at time $t$ can then be approximated as

$$p\big(y_{ij}(t) = 1|D\big) \approx \frac{1}{K}\sum_{k=1}^{K} p\big(y_{ij}(t) = 1|D, \boldsymbol{\theta}^{(k)}\big), \tag{3.5}$$

where $\boldsymbol{\theta}^{(k)}$ is the $k$th sample of $\boldsymbol{\theta}$ from the posterior. The main challenge in this procedure is calculating $p\left(\boldsymbol{t}^{ij}|\boldsymbol{\theta}\right)$, the probability of a dyad's interactions conditional only on the parameters, since doing so requires marginalizing over the the dyad's network path $\{y_{ij}(t) : t \in \boldsymbol{T}^{ij}\}$. This marginalization is intractable for our model due to the continuous-time nature of the Poisson process data and the infinite number of possible network paths.

To circumvent this integration issue, we approximate the probability of each dyad's interactions given the network and parameters by discretizing the underlying network path and restricting potential transitions to a fixed number of times dependent on the observed interactions. Namely, we partition the full time interval $\boldsymbol{T}^{ij}$ based on the data, such that each (sub)interval consists of a single interaction and the surrounding times that are closer to this interaction than any other (see Figure 3.2). We then restrict the network relation to be constant within each of these intervals for the purpose of calculating the probability of the observed interactions. The state of the network edge under this partition is characterized by Markov behavior from the midpoint of one subinterval to the next. The object of inference then changes from a continuous-path $\{y_{ij}(t) : t \in \boldsymbol{T}^{ij}\}$ to the approximation $\left\{y_{ij}(t_l^{ij,*})\right\}_{l=1}^{N_{ij}}$, where $t_l^{ij,*}$ is the midpoint of the interval containing $t_l^{ij}$. To infer the state of the path at other

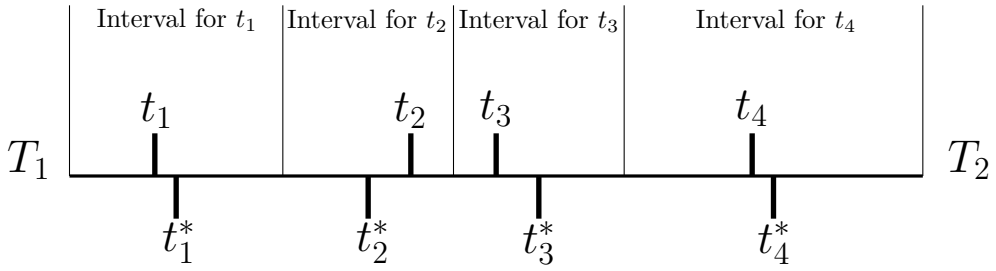


Figure 3.2: Sample partitioning of $(T_1, T_2)$ based on four interactions. Above the horizontal time axis are the time points $\{t_1, t_2, t_3, t_4\}$ at which events were observed. Below the axis are the midpoints of each interval $\{t_1^*, t_2^*, t_3^*, t_4^*\}$. In the interest of parsimony, the network is estimated at each of the midpoint locations and assumed constant within each interval when calculating the probability of the observed data.

times $t$, we simply interpolate between the probabilities of a relation from the nearest two midpoints[1]. Since only one interaction is recorded within each interval and the network is assumed to be slow-changing, we do not expect the underlying network to change repeatedly between interval midpoints.

The procedure we use is not the only possible discretization procedure; for example, one might partition all time intervals $\boldsymbol{T}^{ij}$ into preset, equally spaced intervals. However, one advantage of our chosen method is that it automatically trades off between parsimony and flexibility based on the interaction behavior for a dyad. Tying potential state changes for a network path to interactions focuses modeling effort on periods where there is the most information (arising through interactions) in the data and spends less effort modeling periods of low activity, where there is comparatively little information in the data.

We can view $\left\{y_{ij}\left(t_l^{ij,*}\right)\right\}_{l=1}^{N_{ij}}$ as a hidden Markov model (HMM) with "emissions" $\boldsymbol{t}^{ij}$ arising from a Poisson process with intensity $\lambda_{ij}(t)$. That is, we see evidence of the evolution of the approximated network path, which itself is unobserved, through the relational event data observed in each interval, and borrow strength across intervals to ensure the path is appropriately slow-changing. The "forward-backward algorithm" is a method for estimating

[1]Alternatively, if estimates of the network are desired at certain time points, we can create intervals containing these time points and include them in the approximation.

the hidden states in an HMM (Rabiner, 1989). Using the forward variables from this algorithm, we can marginalize over the latent path to obtain an approximation to the probability $p\left(\boldsymbol{t}^{ij}|\boldsymbol{\theta}\right)$. We can then approximate the posterior $p\left(\boldsymbol{\theta}|D\right)$ in (3.4), and, in turn, obtain posterior samples $\left\{\boldsymbol{\theta}^{(k)}\right\}$ and subsequently estimate $\left\{y_{ij}\left(t\right) : t \in \boldsymbol{T}^{ij}\right\}$ at the midpoints of the intervals $\left\{t_l^{ij,*}\right\}_{l=1}^{N_{ij}}$ using the forward-backward algorithm.

We use Stan (Carpenter et al., 2016) to sample from the posterior distribution of the parameters and latent network. See the Appendices 3.7.2 and 3.8.2 for more details on this procedure.

## *3.4 Proximity Interactions among College Students*

In this section, we analyze interactions between college students from the MIT Social Evolution dataset (Madan et al., 2012), as mentioned in Section 3.1.1. We describe the MIT data in more detail and motivate the model selected. We illustrate one of the advantages of our inferred network over the network snapshots provided by the survey data.

### *3.4.1 The Data*

The data consists of Bluetooth proximity information collected from the phones of 57 MIT undergraduates living in a single dormitory over the period of a school year. Although political leanings and health measures were collected by the surveys, most basic personal information about the students is unavailable (e.g. sex), with the exception of what floor each student lived on and what year of study they were in.

After appropriate data cleaning, as detailed in Appendix 3.7.1, the dataset consists of 66,432 Bluetooth-based proximity interactions among the 1,596 dyads. We model each dyad only over the interval when both students' phones are active (sending/receiving Bluetooth signals) in the proximity logs. Summaries of the data are plotted in Figure 3.3 and reveal strong temporal dependencies. The left panel shows the rate of logged interactions increases throughout the school year. The center and right panels suggests strong daily and weekly periodicity: the majority of interactions occur at night, when students are more likely to be

in their dormitory, and interactivity during the weekend appears to deviate from weekday behavior, with less of a disparity between daytime and nighttime activity levels.

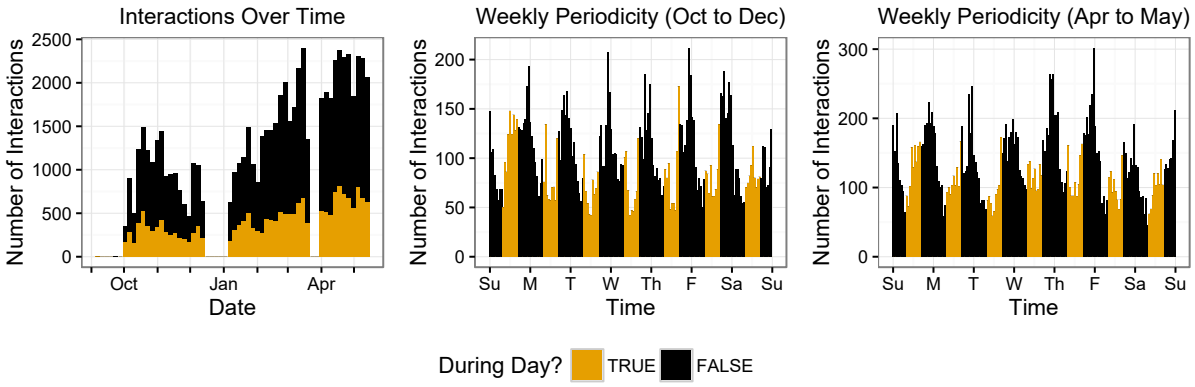


Figure 3.3: Descriptive plots of the MIT Bluetooth proximity interactions. Left: Distribution of interactions over the school year. Center: Interaction frequencies within a week, aggregated from October to December for each dyad. Right: Interaction frequencies within a week, aggregated from April to May.

### *3.4.2 The Model*

Figure 3.3 reveals both weekly periodicity patterns and an increase in activity level as the year progresses. To model the former, we approximate the weekly periodicity with components of its Fourier series representation, choosing the number of sine waves to trade off between simplicity and flexibility. We approximate the trend of increased overall interactivity over the school year by splitting up the year into three terms: first semester (Oct. - Dec.), second semester before spring break (Jan. - March), and second semester after spring break (April - May), and allowing for different activity levels in each period. The similarities between the center and right panels in Figure 3.3 suggest that the pattern of weekly periodicity is largely unrelated to the increase in overall interactivity over the school year.

Dyadic covariates and friendship relations between students are associated with weekly patterns of interactivity and thus are important to explore further in order to build a reasonable model. One unique aspect of the MIT data is that we can leverage the survey data in

model construction. At the level of the dyad, only a weak relation exists between the number of interactions recorded and how often friendship was reported in surveys (see Figure 3.12 in Appendix 3.7.1). However, when aggregating across dyads, we find significant differences in the interactivity patterns between reported friends and non-friends. Using survey data as a proxy for friendship, we compare the weekly patterns for reported friends/non-friends and individuals who live on the same/different floors, where we consider a dyad to be friends if at least one student named the other as a "close friend". In the right panel of Figure 3.4, non-

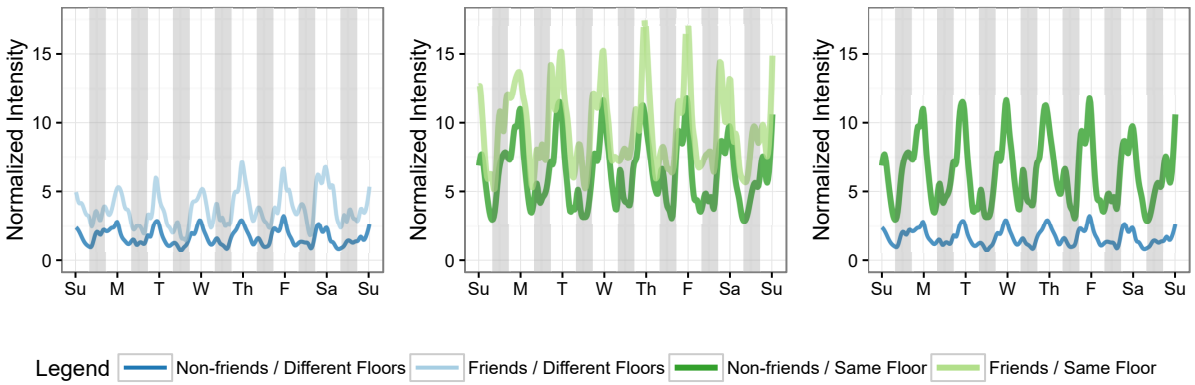


Figure 3.4: A comparison of weekly interactivity patterns by survey-reported friendship and whether or not pairs of individuals reside on the same floor. Daytime is highlighted with grey bars. Intensities are normalized so that non-friends on different floors have mean intensity of one. In order to avoid potential friendship dynamics in this comparison, we consider a pair of students "friends" if at least one of the two identified themselves as "close friends" in four of the five surveys conducted. We consider a pair of students "non-friends" if they jointly identified themselves as "close friends" once or less.

friends that live on the same floor tend to interact with higher propensity than non-friends that on different floors. However, while their overall interactivity levels differ, individuals in these two groups interact in similar weekly patterns. The left and center panels suggest that friends interact more than non-friends regardless of floor status, although the relative compounding effect is stronger for friends who live on different floors. Additionally, we find some evidence that friends may interact more than non-friends during the daytime (defined here to be 8 AM to 5 PM). Since the students all live in the same dormitory, encounters

during the day, when more conscious effort is required to be in close proximity to others, may be especially indicative of friendship.

Using the notation in (3.2), we construct the following interaction model:

$$w\left(t, \text{floor}(i), \text{floor}(j)\right) = c_{0,\text{term}}\left(1 + c_1 \mathbf{1}_{\text{floor}(i)=\text{floor}(j)}\right)\left(k_0 + \sum_{l=1}^{3} k_{l,1}\cos\left(k_{l,2}t + k_{l,3}\right)\right) \quad (3.6)$$

$$m\left(\text{floor}(i), \text{floor}(j)\right) = c_{2,1}\mathbf{1}_{\text{floor}(i)\neq\text{floor}(j)} + c_{2,2}\mathbf{1}_{\text{floor}(i)=\text{floor}(j)} \quad (3.7)$$

$$a\left(t, \text{floor}(i), \text{floor}(j)\right) = c_{0,\text{term}}\left(1 + c_1 \mathbf{1}_{\text{floor}(i)=\text{floor}(j)}\right) c_3 \mathbf{1}_{\text{daytime}}\left(t\right) \quad (3.8)$$

In the baseline $w$, three sets of sinusoidal terms model the weekly behavior while multipliers for school term and floor govern the overall levels of interactivity. We characterize friendship by an increase in daytime activity relative to the baseline pattern of weekly periodicity and an overarching multiplicative increase in intensity, dependent on floor status. Based on Figures 3.3 and 3.4, we assume this increased propensity to interact during the day for friends is subject to the same term- and floor-dependent modifiers that characterize general interactivity between dyads and use the same multipliers $c_{0,\text{term}}$ and $c_1$ in the daytime adjustment (3.8) that govern the overall interactivity in the baseline (3.6).

We can similarly use the survey data to inform our model for the network paths, treating the reported relationships as observations of CTMCs at the survey dates. In Figure 3.5, living on the same floor and being in the same year have independent, multiplicative effects on the empirical probability of friendship for dyads. Additionally, these proportions are constant throughout the school year and changes between survey dates are consistent with a single, low transition rate $q$. We model the network paths as CTMCs with sparsity

$$s\left(\text{floor}(i), \text{floor}(j), \text{year}(i), \text{year}(j)\right) = \left(1 + s_1 \mathbf{1}_{\text{floor}(i)=\text{floor}(j)}\right)\left(1 + s_2 \mathbf{1}_{\text{year}(i)=\text{year}(j)}\right) s_0 \quad (3.9)$$

and constant transition rate $q$. Introducing this structure into the sparsity model encourages block model behavior in the underlying network, where students who live on the same floor or

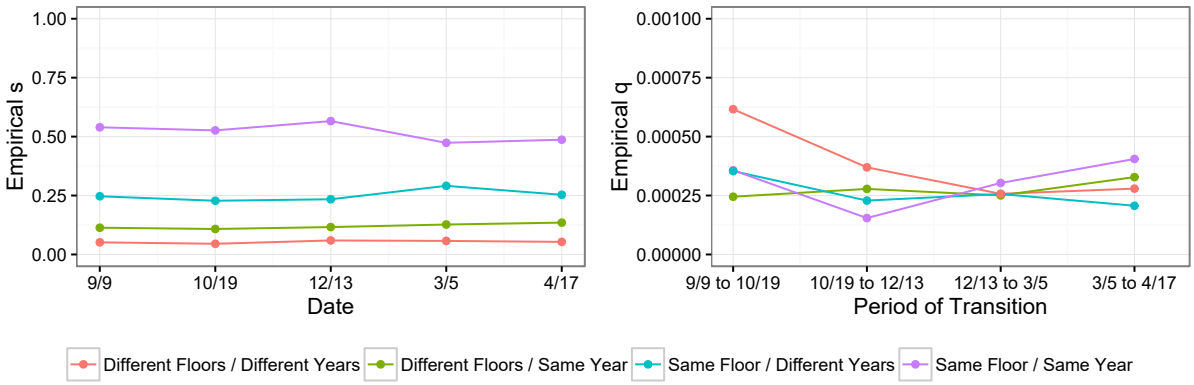


Figure 3.5: Empirical CTMC parameter values for the four possible dyadic floor/year pairings derived from survey data. Sparsity $s$ is estimated by the proportion of dyads who report friendship at each date, while transition parameter $q$ is derived from the proportion of dyads which change reported friendship between consecutive survey dates.

are in the same year are more likely to be friends with one another and hence form "blocks" of friendship.

Before we can obtain estimates of the student network, we must assign priors to the parameters of our model. We observe a strong cyclical pattern in communications (associated with, for example, day and night), so we fix $k_{i,2}$ using a Fourier transform on the observed weekly pattern of interactivity. We set the interactivity multiplier for the first school term equal to one ($c_{0,\mathrm{t1}} = 1$) to ensure the identifiability of these multipliers. We specify weakly informative priors on the remaining parameters based on support matching; additional details on these prior distributions are presented in Appendix 3.7.2.

### *3.4.3 Network Analysis*

We used Stan to generate 9,000 samples (after burn-in) of the parameters from the posterior distribution. Details of the sampling procedure are provided in Appendix 3.7.2.

One of the key benefits of the inferred network is that it evolves in continuous-time, allowing for a more detailed examination of network evolution. For example, in the left two panels of Figure 3.6 we plot two snapshots of the MIT network as self-reported by students.

In the first survey date shown, December 2008, student #2 (highlighted, towards the top

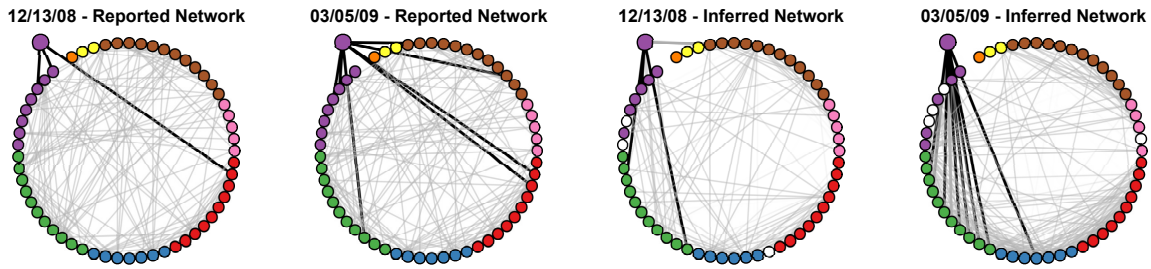


Figure 3.6: Snapshots of networks of the MIT students at two survey dates. Students are colored by the floor they lived on. For the reported network, students are considered friends if either reports the other as a "close friend." In the case of the inferred networks, students are colored white if their phones are inactive and we can no longer infer relations for them. Uncertainty in the inferred relations is quantified by the opacity of the ties. In all snapshots, student #2 is highlighted, as well as his/her relations.

left) was reported to be friends with three other individuals. By March of the following year, there were eight friendships involving this student. Between the two dates, six friendships were formed and one dissolved. Yet, with only this information, we have little insight into how these changes came about and cannot test theories about the social processes underlying these changes, which are of particular interest for actor-oriented models (Snijders, 1996).

The continuous-time evolution of our inferred network provides greater insight into potential social theories, as we can observe snapshots of the network at additional time points. In the right two plots of Figure 3.6 we present snapshots of the inferred network at the same two survey dates. We observe similar behavior with student #2, albeit with different individuals; we inferred three friendships in December and seven in March, corresponding to five formed friendships and one dissolved. Since the inferred network evolves in continuous-time, we can examine snapshots of the network at any time in between the survey dates to get a better sense of how the network evolves and see how the interactions influence the formation/dissolution of inferred ties.

In Figure 3.7, we plot a subset of the inferred network for six additional dates between the survey dates. An animated series of all daily snapshots between the survey dates can

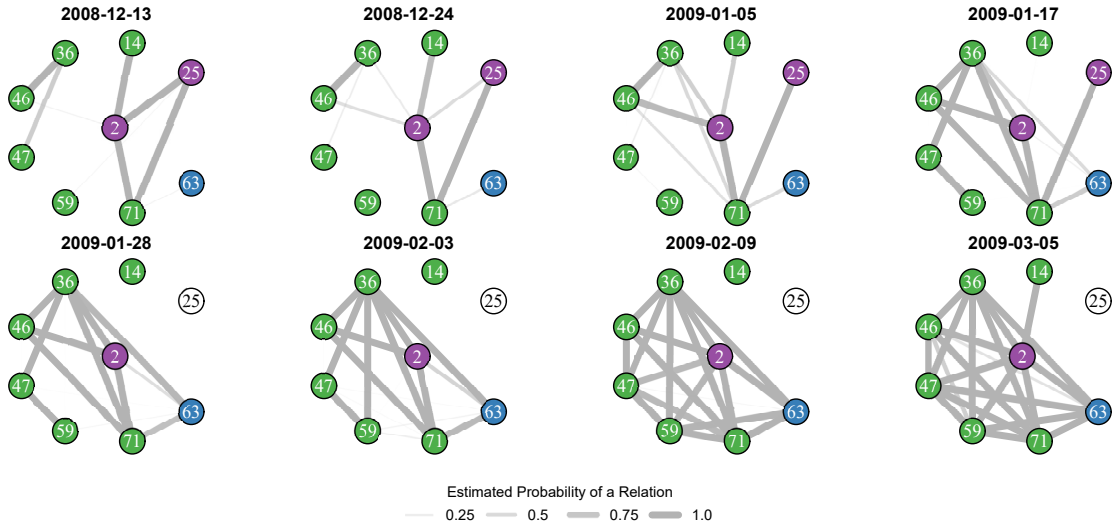


Figure 3.7: Snapshots of the inferred network between December 13th, 2008 and March 5th, 2009 for a selected subset of students. Students are colored by the floor they lived on. Edge widths for each dyad denote the estimated mean probability of friendship at the given time.

be found at `https://git.io/viTtS`. These intermediate snapshots of the network provide evidence of triadic closure in this community, as friendships do not tend to form randomly but rather appear with higher propensity between actors who have mutual connections. We would have missed this behavior without the ability to examine the evolution of the network at finer resolutions; we were able to find evidence of this phenomenon only after examining intermediate snapshots of the network.

## *3.5 Proximity Interactions among Barn Swallows*

We now turn to the behavior of North American barn swallows (*Hirundo rustica erythrogaster*). Unlike in the MIT student data, there is no survey data to guide our selection of the functions $w$, $m$, $a$, $s$, and $q$ underpinning the model.

### *3.5.1 The Data*

Telemetry data were gathered for 17 barn swallows monitored from 06:00-09:00 and 17:00-20:00 (periods of high barn swallow activity) over three consecutive days during mating

season (Levin et al., 2015, 2016). Barn swallows were outfitted with tracking devices which logged the presence of nearby swallows every 20 seconds. After data cleaning, detailed in Appendix 3.8.1, the resulting dataset has 1009 interactions among its 136 dyads. Of the barn swallows studied, 10 were male and 7 were female. Various other physical characteristics of the birds are also available, including mass, tail streamer length, and plumage color.

### *3.5.2 The Model*

We choose to omit various observable covariates known to be associated with barn swallow social behavior from our model. For example, our model does not explicitly encode for factors that influence mate selection, such as the female preference to mate with males with darker plumage color (Safran and McGraw, 2004; Safran et al., 2005). Instead, our strategy is to develop a parsimonious model of the marginal network behavior without considering observables endogenous to network formation. This strategy, which may be necessary for less well-studied communities, provides an overall characterization of barn swallow social behavior that is useful in its own right and, as we show in the results, can also be correlated with observable features after model fitting.

In Figure 3.8, we see that both sex and time play important roles in understanding interaction behavior between swallows. The left panel shows that female birds interact very little with one another, yet interact quite often with males. In contrast, male birds regularly interact with each other. We take these differences in activity levels to represent behavioral discrepancies in the types of relationships formed by the various sex pairings rather than differences in the propensity for forming these relationships themselves. The right panel shows the rates of interaction during the periods of observation are non-constant. Differences in these rates may be due to factors outside of our interest, such as the weather and other variable environmental stimuli, which are unlikely to impact network dynamics.

Under the assumption that swallows without social ties interact at a rate independent of

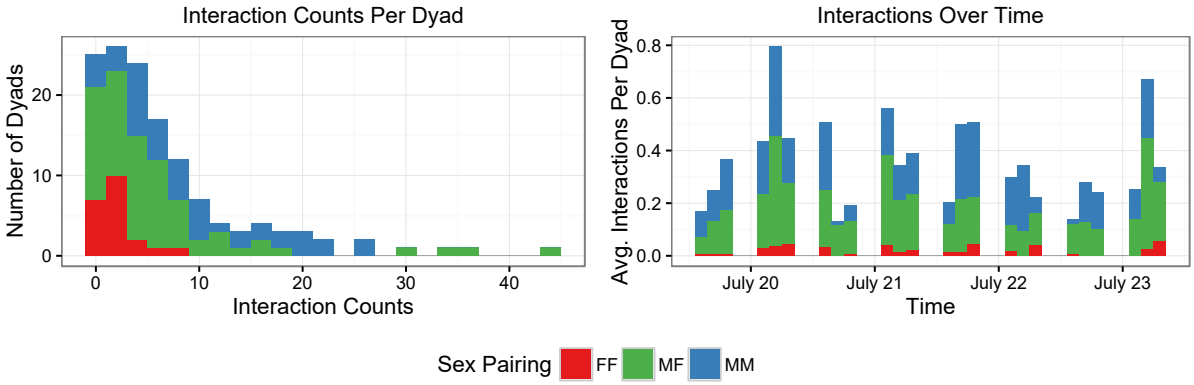


Figure 3.8: Descriptive statistics of the interaction data. Left: Histogram of the number of interactions for each dyad, color coded by the sex pairing of each dyad. Right: Histogram of the number of interactions over time, aggregated across dyads. Interactions are also color coded by its dyads' sex pairing.

their sex pairing, we choose to model the baseline activity rate for any pair at time $t$ as

$$w(t) = k_1 \mathbf{1}_{7/19 \text{ PM}} + k_2 \mathbf{1}_{7/20 \text{ AM}} + ... + k_8 \mathbf{1}_{7/23 \text{ AM}}. \tag{3.10}$$

In this model, baseline activity rate is constant within each observational period but varies from period to period. This allows for flexibility in the temporal trend without imposing a strict parametric trend.

In contrast, we believe that barn swallows have different types of social relationships depending on the sexes of the birds involved and thus sex pairing will affect the activity levels of familiar birds. We embed this phenomenon by allowing for different multiplicative increases in interaction rates depending on sex pairing. Omitting an additive effect $a$ to the baseline intensity, we assume that birds with social ties interact with the same temporal patterns as birds without ties. The intensity rate of the $(i, j)$th bird pair at time $t$ is then given by

$$\lambda_{ij}(t) = \left(1 + \left(c_{FF}\mathbf{1}_{FF} + c_{MF}\mathbf{1}_{MF} + c_{MM}\mathbf{1}_{MM}\right) y_{ij}(t)\right) w(t), \tag{3.11}$$

where $y_{ij}(t)$ is an indicator for whether or not $(i, j)$ are socially connected at time $t$.

In accordance with our strategy to not explicitly encode observables into the formulation of the network, we model each dyad's connection $y_{ij}(t)$ with i.i.d. CTMCs with common sparsity $s$ and common transition parameter $q$. We complete the model by assigning weakly informative priors based on support matching to the parameters $\{k_1, ..., k_8, c_{FF}, c_{MF}, c_{MM}, s, q\}$. Detailed priors are given in Appendix 3.8.1.

### *3.5.3 Network Analysis*

We considered 9,000 samples (after burn-in) obtained from Stan (see Appendix 3.8.1 for computational details). A full animation of our inferred network, alongside the raw encounter network, is available at `https://git.io/viTtP`.

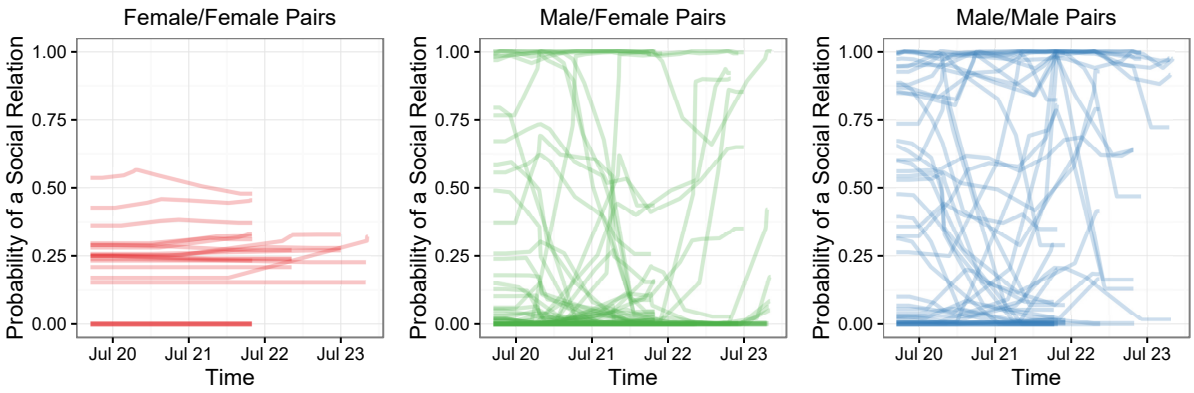


Figure 3.9: Estimated mean paths grouped by sex pairing.

In Figure 3.9, we group the estimated mean network paths by sex pairing. In the left panel, we see the estimated social connections between female birds are generally static, without large changes in the probability of social ties over the course of the study. Unfortunately, some of the tracking devices on the female swallows became inactive early in the study, obscuring potential changes in the latter half of the study.

Displayed in the right panel of Figure 3.9, pairs of male swallows exhibited a very diverse set of behaviors. In stark contrast to pairs of female swallows, many male pairs had dynamic associations in which their interactivity changed dramatically over the time of the study. The

probability of a social tie fluctuated by at least 50% over the three day period for slightly more than 20% of possible male pairs. The differences in behavior may be attributable to the aggression of males during mating season, when they are particularly prone to engaging in territorial conflict with one another. At the same time, other pairs had relatively stable social ties or lack thereof. For example, we estimated 10 out of the possible 45 pairs to be socially connected with probability 0.8 or higher when averaged over the course the study. These relations are plotted in Figure 3.10. We observe a clique of four male swallows, as well

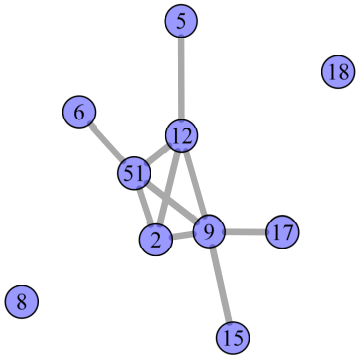


Figure 3.10: Network of male swallows. Nodes are labeled with each bird's tag identification number. Edges denote "strong" relationships, between birds who are predicted to be socially connected with probability 0.8 or higher on average over the study.

as four other males with fewer relations to the clique. The stable behavior between these males may be evidence of extended territorial conflict or more complex social structure, such as unmated males "attending" nests of other birds in the colony (Crook and Shields, 1987).

Both strong stability and some dynamics characterize the behavior of male/female pairs, displayed in the middle panel of Figures 3.9. We consistently classify about 75% of the possible male/female pairings to have no relation. The remaining 25% displayed a wide range of relational probabilities and dynamics. Among the latter were four disjoint pairs which averaged probability of relationships of 0.8 or higher. While barn swallows are sexually polygamous, they are socially monogamous and form pair-bonds for the breeding season; the behavior of these four pairs suggest that they may be such pairs.

Previous studies of European barn swallows established that females prefer males with longer tail streamers (Moller et al., 1998), while in the United States it has been documented that females prefer males with darker ventral plumage (Safran et al., 2005). Despite not encoding sexual selection into our model, we find evidence congruent with both theories for the studied North American barn swallows. Under the assumption that male/female relationships signal mating preferences, the males which were most successful in establishing bonds with females tended to have longer tail streamers and darker ventral plumage. Comparing the nine males for which tail streamer length was recorded, the three males with highest probability of relations with females had the three longest tails (among the nine). This finding, illustrated in Figure 3.11, is quite robust to the choice of cutoff dividing males by tail streamer length. A similar relationship held to a lesser extent for ventral coloring, which was strongly correlated ($r$=0.79) with tail streamer length for the studied males.

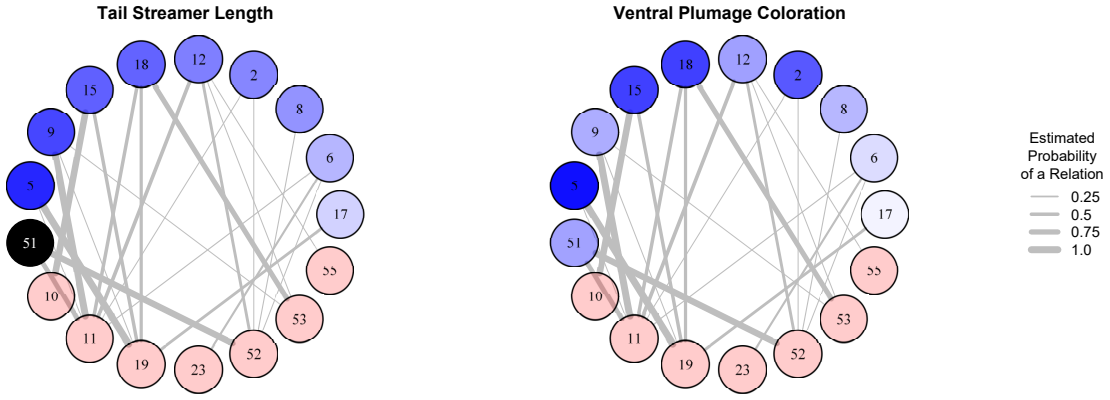


Figure 3.11: Male/female relationships among barn swallows. Red nodes denote females, while blue and black nodes denote males. Edges are weighted by average probability of a social relation, and estimated probabilities lower than 0.1 are omitted. In the left graph, males are shaded by tail streamer length, with darker color corresponding to longer tails (except for black, which corresponds to a missing value). In the right graph, males are shaded according to the color of their ventral plumage, with darker blue corresponding to darker coloration.

## *3.6 Discussion*

The proliferation of continuous-time relational event data presents both new opportunities and challenges for social network research. Event data represent observed behavior, which fundamentally differs from reported behavior, the traditional domain of survey data (Killworth and Bernard, 1976; Eagle et al., 2009). Many of the network analysis tools developed for survey data are not appropriate for relational event data (Howison et al., 2011), and new methods must be proposed. Simultaneously, numerous ethical issues should be raised involving how the data should be used and how to ensure privacy of individuals.

We have presented a hierarchical network model in which relational events are modeled with Poisson processes governed by dyadic covariates and a dynamic latent network. Our model acknowledges the intrinsic nature of relational event data as continuous-time interactions only indirectly indicative of underlying dynamic network relations. Since a network of relational ties is often the primary object of interest for researchers, models which address the discrepancy between the observed activity data and these relational ties are important in order to use these proliferating sources of data.

Depending on the intended application, further developments of the model may allow for more precise estimates of the latent network or more fully capturing dependence between events. For example, in organizational email data, emails can have multiple recipients, deviating from our dyadic framework, and each email is associated with content that informs the nature of the interaction. Differentiating between these types of interactions through the use of multiple connected point processes may yield more accurate estimates of our network of interest. Alternatively, we may want our model to discriminate between the various types of relationships between workers through the use of multiple latent networks. Economic phenomena, such as crisis management and information diffusion, may occur through different channels of interest. As another example, when modeling botnets, mutually-exciting Hawkes processes may be more suitable than independent Poisson processes for modeling the spread of malware, since malware can only be transmitted from devices that are already themselves

infected. The structure of relational event data, its relationships to an underlying network, and the meaning of the network will vary considerably across applications. Hence, domain knowledge should be used whenever possible to customize the model to target capturing relevant and interesting structure.

## *3.7 Appendix: MIT Social Evolution Data*

### *3.7.1 Data Considerations*

Participants in the study were given smartphones outfitted with an application that used Bluetooth to scan for other nearby (within 10 meters) smartphones approximately every six minutes. However, the application only logged the presence of other devices set as "discoverable." Due to this and other technical issues, many of the reported encounters were non-reciprocal and it is possible many encounters went unreported. In addition, a fraction of the logged encounters occurred between phones on different floors due to vertical proximity. These encounters are not of interest, and are identifiable as being estimated by the application to have low probability of being on the same floor using WLAN location data. We chose to drop all encounters with probability 0.2 or lower. In addition, we only considered interactions that occur during the Fall (September 3rd to December 13th) and Spring sessions (January 5th to March 20th and March 30th to May 15th), notably excluding winter and spring breaks, which were extended periods when school is not in session. Please refer to Madan et al. (2012) for further details about the data.

We converted the proximity logs into undirected (reciprocal) interaction data by combining the directed proximity logs for pairs of individuals, taking encounters logged in either phone. Logs that were time-stamped within 30 minutes of one another were taken to correspond to the same interaction.

Defining a dyad to have reported as friends in a survey if either marks the other as a 'close friend,' we observed significant dissonance between the survey data and interaction data on

a dyadic level. Over twenty percent of dyads who reported friendship in all five surveys did not have any recorded interactions, compared to just under fifty percent for dyads who never reported friendship. The weak relation between the sources of data was robust to different choices of defining friendship in the survey data and raises validity concerns about both sets of data.

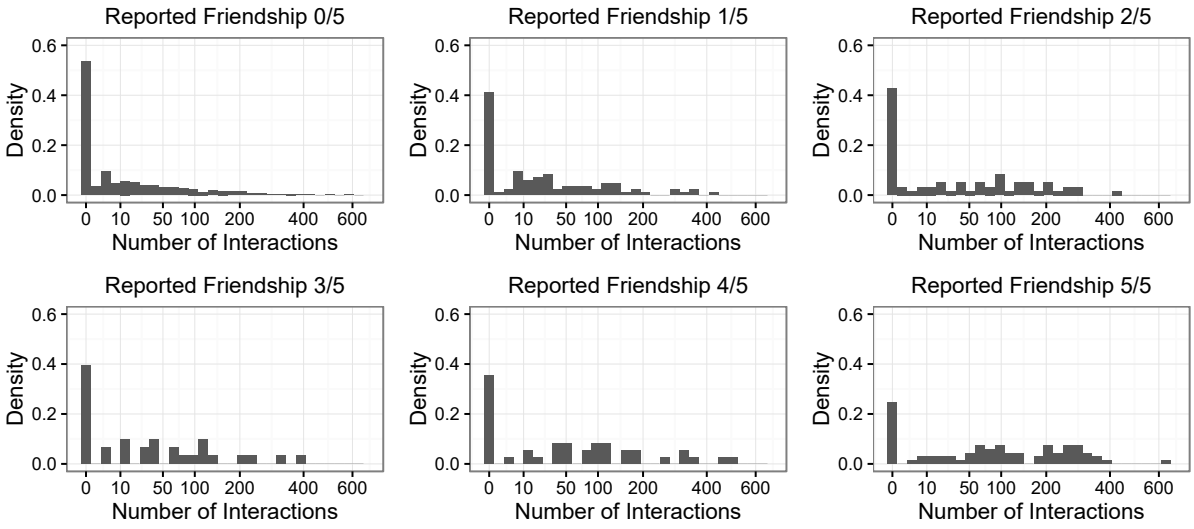


Figure 3.12: Distributions of the total number of recorded interactions per dyad, where the data is partitioned by the fraction of times each dyad reported friendship in the five surveys. We consider a dyad to have reported as friends in a survey if either individuals marks the other as a 'close friend.'

### *3.7.2 Priors and Model Fitting*

We assign the following independent priors to the parameters of the model specified by (3.6)-(3.9)): (where applicable, units are in hours)

$$\begin{aligned}
\{c_{0,\mathrm{t}2}, c_{0,\mathrm{t}3}\} &\sim \mathrm{Exp}\,(1) & c_1 &\sim \mathrm{Exp}\,(1)\\
\{c_{2,1}, c_{2,2}\} &\sim \mathrm{Exp}\,(100) & c_3 &\sim \mathrm{Exp}\,(1)\\
k_0 &\sim \mathrm{Exp}\,(1) & \{k_{i,1} : i \in \{1,2,3\}\} &\sim \mathrm{Exp}\,(1)\\
\{k_{i,3} : i \in \{1,2,3\}\} &\sim \mathrm{Uniform}\,(0, 2\pi) & s_0 &\sim \mathrm{Beta}\,(1, 49)\\
s_1 &\sim \mathrm{Exp}\,(1) & s_2 &\sim \mathrm{Exp}\,(1)\\
q &\sim \mathrm{Exp}\,(1e10)
\end{aligned}$$

We used Stan to generate 10,000 samples using a single MCMC chain, initializing each of the parameters at its prior mean and discarding the first 1,000 as burn-in. This sampling procedure took approximately 24 hours on a standard laptop. We assessed convergence by visually examining the trace plots shown in Figure 3.13 and through estimates of effective sample size calculated via Stan, which were greater than 3,000 for all estimated parameters.

## *3.8 Appendix: Barn Swallow Data*

### *3.8.1 Data Considerations*

Every 20 seconds, the proximity loggers attached to each barn swallow logged the presence of other barn swallows within 5 meters. This threshold was chosen by the researchers since it represented the average distance between nests at the observation site. The devices split logged encounters between birds that lasted more than 5 minutes into 5 minute intervals. Unfortunately, the proximity loggers were imperfect; sometimes encounters recorded by one bird were not recorded by the other and the batteries of some devices died (starting on July 22nd) before the end of the observation period. See Levin et al. (2015) for a more comprehensive description of the data.

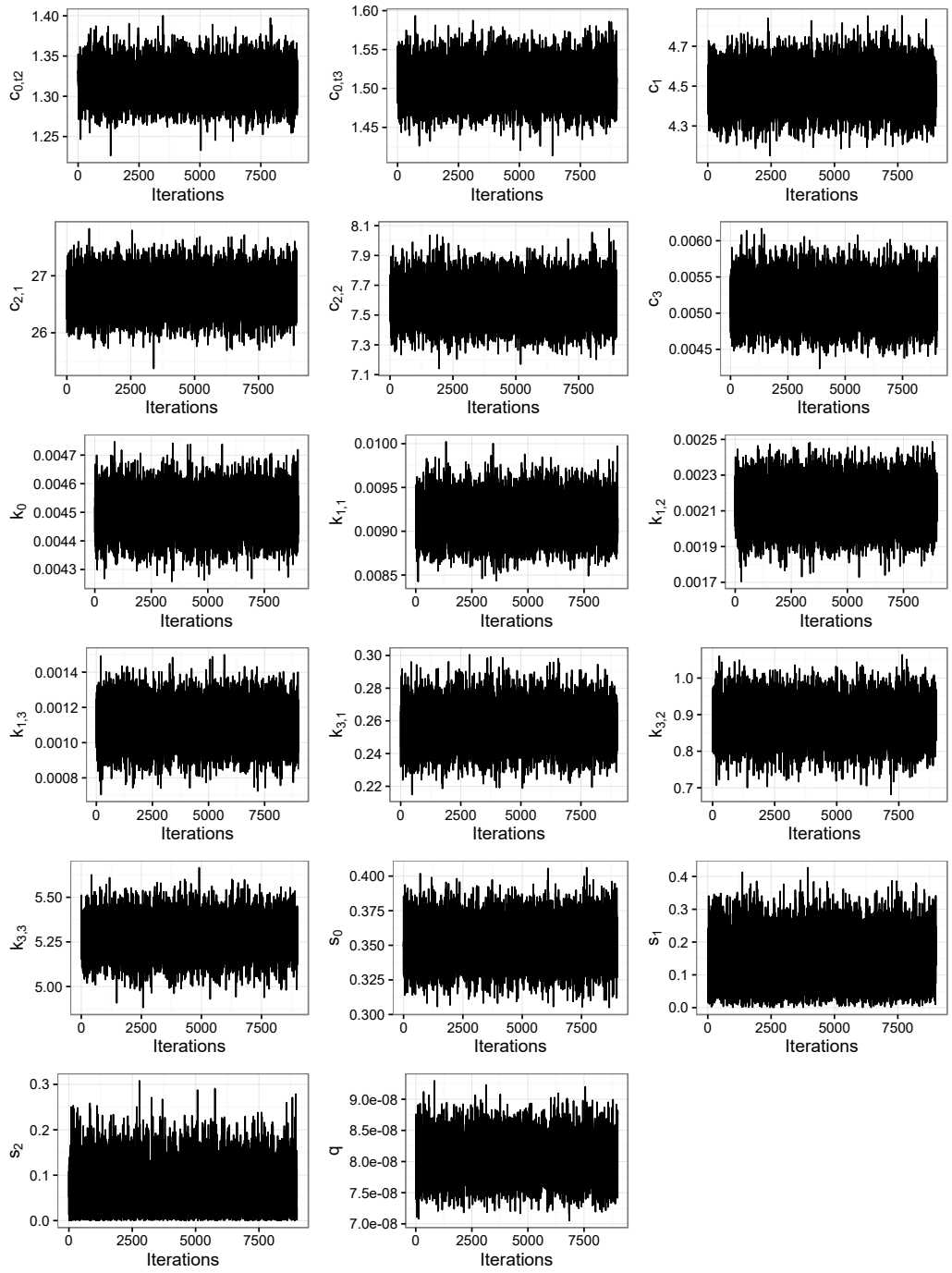


Figure 3.13: Trace plots for the MIT model parameters. Values obtained during the burn-in period are discarded.

We accounted for asymmetries and extended encounters in the encounter log. To ensure encounters that last longer than 5 minutes were treated as a single interaction, encounters separated by thirty seconds or less were merged. To compensate for the imperfect behavior of the tracking devices, we combined the logged encounter history for every pair of birds, taking encounters logged in either birds' tracker as an interaction and merging overlapping encounters. Lastly, to help avoid issues related to battery-life, we only modelled each dyad over the interval when the trackers of both birds recorded activity.

### *3.8.2 Priors and Model Fitting*

We assign the following priors to the parameters of our model: (units for $k_i$ and $q$ are provided in hours)

$$\{c_{FF}, c_{MF}, c_{MM}\} \sim \text{Exp}\,(2) \qquad \{k_i : i \in \{1, ..., 8\}\} \sim \text{Exp}\,(1)$$

$$s \sim \text{Beta}\,(1, 9) \qquad q \sim \text{Exp}\,(1000)$$

Using Stan, we ran a single MCMC chain for 10,000 iterations and discarded the first 1,000 as burn-in. The parameters were initialized at the means of their respective prior distributions, and the entire sampling procedure took approximately 6 minutes. As with the previous application, we assessed convergence by examining trace plots, provided in Figure 3.14.

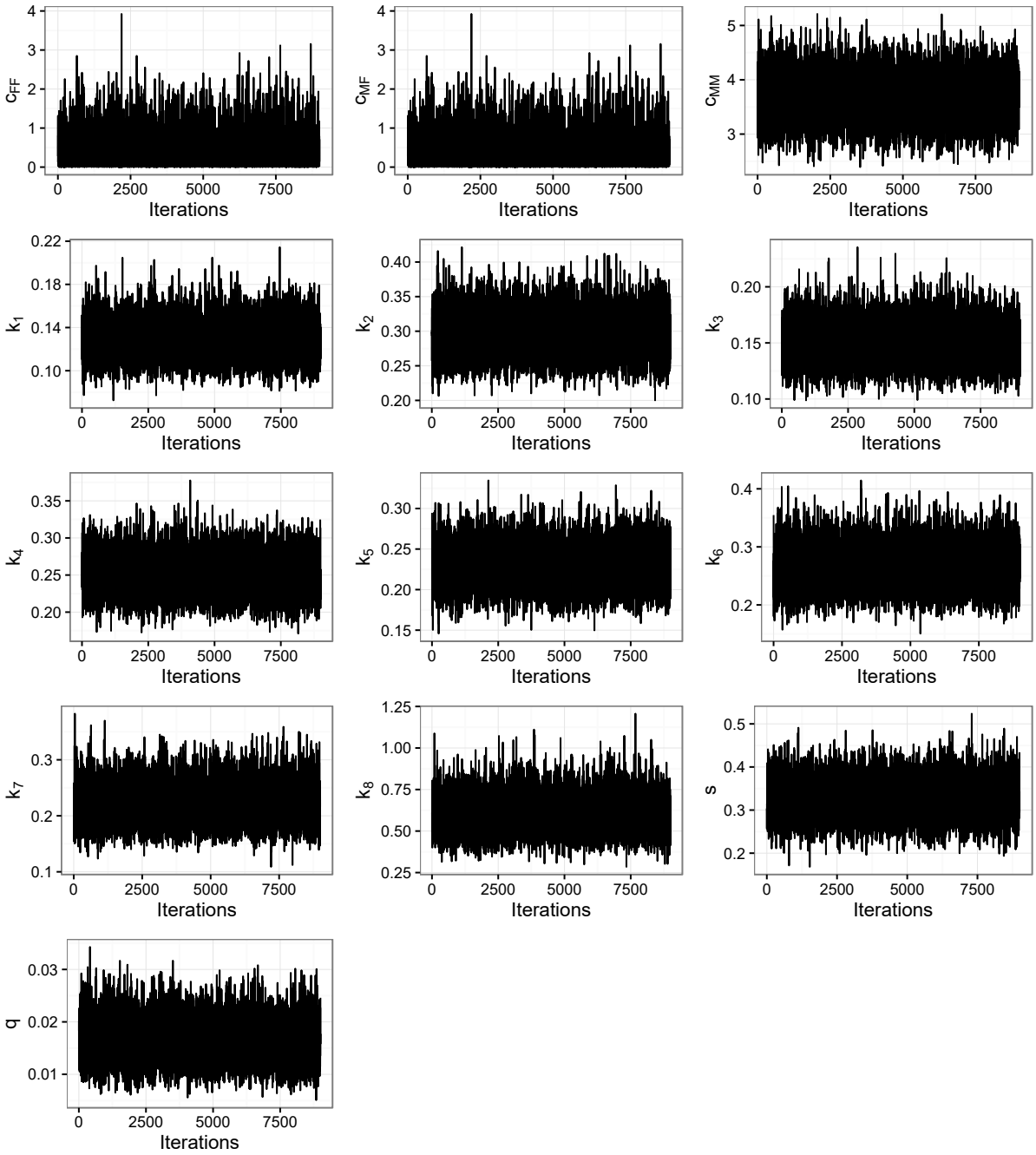


Figure 3.14: Trace plots for barn swallow model parameters. Values obtained during the burn-in period are discarded.

# Chapter 4

# ANOMALY DETECTION IN LARGE SCALE NETWORKS WITH LATENT FACTOR MODELS

Joint work with Tyler H. McCormick.

## *4.1 Introduction*

The proliferation of relational event data has allowed for network data to be collected on increasingly larger networks. Increases in $N$, the number of subjects or nodes in a network, are accompanied by corresponding increases in the number of directed edges, or dyads, to $N^2$. In large networks, the $N^2$ number of dyads under consideration generally precludes the use of statistical models with dyad-specific parameters, such as that of our model in Chapter 3 which individually quantifies each dyadic relation. On the other hand, latent space models (Hoff et al., 2002) are relatively parsimonious and represent dyadic relations using node-specific parameters. A strength of these models is their ability to share information across dyads involving the same node, thus allowing predictions on dyads even with missing or otherwise unavailable data. While initially developed for static networks, these models have been extended to incorporate discrete-time network dynamics (Sewell and Chen, 2015; Durante and Dunson, 2016). As with most network models, estimation, primarily done through Markov chain Monte Carlo (MCMC), can be computationally intensive as the likelihood involves $N^2$ terms, one for each dyad.

One particular setting where these computational considerations are particularly pertinent is cybersecurity-related anomaly detection on computer networks. As a motivating dataset, consider Netflow activity data collected by the Los Alamos National Laboratory

(LANL) on their enterprise network for a period of 89 days (Turcotte et al., 2018)[1]. Each record consists of directed communication between network devices, and network activity is logged between over 25,000 computers over these 89 days. In cybersecurity, a major objective is to flag potential intrusions on the network. The detection of these invaders is time sensitive, reflecting a desire to prevent further intrusion when possible.

The typical approach for anomaly detection in this setting is to build a model that represents normal behavior and use deviations from this model to flag potential anomalies (Neil et al., 2013a). The strength of a underlying dyadic relation would then correspond to the propensity for that dyad to interact. These models tend to be relatively simple, as estimation must be implementable in a fully online algorithm. Namely, algorithms which require evaluating the full likelihood via summation of over 600 million dyads are infeasible. Existing models generally avoid complex interaction terms, opting for simple sender- and receiver-specific popularity terms (Neil et al., 2013b) or clustering nodes and modeling interactions at a cluster level (Metelli and Heard, 2016).

In this chapter, we consider modeling relational event data for large, sparse networks. We model activity as arising from a network characterized by a dynamic latent factor model or bilinear mixed-effects model from Hoff (2005), which represents an direct increase in complexity from the model of Neil et al. (2013b) by including a bilinear interaction term for sender- and receiver-specific latent factors. In contrast to Chapter 3, we focus on the issue of scalability with respect to estimating the underlying network and not the continuous time nature of the event data, restricting instead the network to evolve with discrete-time dynamics. Estimation is an application of the variational message passing algorithm described in Minka (2005) with a case control approximation inspired by Raftery et al. (2012) in order to reduce the computation cost from $O(N^2)$ to $O(E)$, where $E$ is the number of observed edges in the network. Taking advantage of the parametric form of the variational approximation to the posterior, we allow our parameters to update with discrete time dynamics via Gaussian

[1]The data is available at `https://csr.lanl.gov/data/2017.html`.

random walks, adopting the autotuning procedure from McCormick et al. (2012) in order to flexibly adjust the amount of additional variation introduced at each time step.

The closest existing work to our own are that of McCormick et al. (2012) and Salter-Townshend and Murphy (2013), both of which take variational approaches to other logistic regression models. In addition to their autotuning procedure, McCormick et al. (2012) propose a general purpose algorithm for estimating dynamic logistic regression models. However, their approach jointly updates the logistic parameters via Newton's method and requires inversion of the corresponding Hessian. When the number of parameters in the model is large, such as when each node has a specific popularity term, this process becomes infeasible. On the other hand, Salter-Townshend and Murphy (2013) focus on a variational algorithm for the static latent position model (Hoff et al., 2002) based on minimizing KL divergence, finding substantial computational gains over a comparable MCMC even before using the case-control approximation from Raftery et al. (2012). The primary expectation required in the algorithm is inherently intractable, and they proceed via a series of Taylor series expansions in order to reach an tractable expression. In contrast, we choose to adopt a variational approach minimizing a different divergence metric but resulting in a tractable, analytic set of updating equationns.

The chapter is organized as follows. In Section 4.2.1 we describe the static bilinear effects model. In Section 4.2.2 we present the variational message passing algorithm for the static model as well as the case-control modification. Section 4.2.4 adapts the model and algorithm to a dynamic setting, and Section 4.2.5 describes anomaly detection after estimation is complete. In Section 4.3 we demonstrate the performance of our algorithm in a simulation study, in Section 4.4 we apply our algorithm to the LANL computer network data, and in Section 4.5 we conclude.

## 4.2 Dynamic Latent Space Models

### *4.2.1 Bilinear Mixed-Effects Model*

In this chapter we focus on the logistic specification of the bilinear mixed-effects model (Hoff, 2005). Letting $y_{i,j}$ indicate the presence of directed activity from sender $i$ to receiver $j$, under this model

$$y_{i,j} = \text{Bernoulli}(p_{i,j}), \tag{4.1}$$

where

$$\text{logit}(p_{i,j}) = \mu + \alpha_i + \beta_j + u_i^T v_j. \tag{4.2}$$

$\mu$ controls the overall sparsity of the network, while $\alpha_i$ and $\beta_j$ represent sender- and receiver-specific popularity terms and $u_i$ and $v_j$ are $d$-dimensional sender and receiver-specific latent factors. The interaction term $u_i^T v_j$ captures the affinity between $i$ and $j$, and can be interpreted as the additional propensity for senders with certain latent characteristics to interact with receivers with other certain latent characteristics over the baseline propensity implied by their respective popularities.

We complete our Bayesian model by introducing independent Gaussian priors for each of the parameters:

$$\begin{aligned}
\mu &\sim N(0, \sigma_\mu), \\
\alpha_i &\sim N(0, \sigma_\alpha), \\
\beta_j &\sim N(0, \sigma_\beta), \\
u_i &\sim N(0, \Sigma_u), \\
v_j &\sim N(0, \Sigma_v).
\end{aligned}$$

Lastly, we use $N$ to denote the number of nodes in the network and the $N \times N$ matrix $\mathbf{Y}$ to denote all directed activity in the network. Without loss of generality, we assume the number of senders and the number of receivers in the network are equal.

### *4.2.2 Variational Inference*

Let $\theta \equiv \{\mu, \alpha_i, ...\beta_j, ...u_i, ...v_j...\}$ denote the set of latent variables in the bilinear mixed-effects model. Given the large number of parameters, we wish to construct a parsimonious representation of the posterior $p(\theta|\mathbf{Y})$. For example, the posterior covariance between the sender- and receiver-specific popularity terms would require the storage of a $N \times N$ matrix. To this end, we focus on learning a fully-factorized approximation $q(\theta)$ to the posterior, with independent terms for each latent variable. Specifically,

$$q(\theta) = q(\mu) \prod_i q(\alpha_i) q(u_i) \prod_j q(\beta_j) q(v_j), \tag{4.3}$$

where each marginal term is modeled with a Gaussian (with $d \times d$ covariance matrices for each of the latent factor terms). Representing the posterior of each latent variable with an independent Gaussian leads to a storage complexity of $O(N)$ and will also help facilitate the introduction of temporal dynamics in the following section.

Inference proceeds as a direct application of power expectation propagation (Power EP) (Minka, 2005) and takes the form of a message passing algorithm. We describe the algorithm in detail below but omit some of the theoretical basis provided in (Minka, 2005).

First, we can recast the posterior as a product of factors, where each factor is either an dyadic observation or a prior over a latent variable.

$$p(\theta|\mathbf{Y}) \propto p(\mathbf{Y}|\theta)\, p(\theta) \tag{4.4}$$

$$\propto \prod_{(i,j)} f_{(i,j)}(\theta) \tag{4.5}$$

We use $(i, j)$ to index the factor denoting the directed dyad $i \to j$, with $(0, k)$ for the factor denoting the prior over the $k$th latent variable $\theta_k$ [2]. We cast our fully-factorized approximation $q$ (see equation (4.3)) in light of the same factors such that the approximation to the

[2] We arbitrarily index the set of latent variables in our model with $k$ for notational simplicity for contexts in which the differences between the various latent variables are unimportant.

posterior for $\theta_k$ is the product of messages from each of the factors to $\theta_k$:

$$q(\theta_k) = \prod_{(i,j)} m_{(i,j)\to\theta_k}(\theta_k). \tag{4.6}$$

Rearranging terms, we can also view the product of messages from a factor $(i,j)$ as an approximation $\tilde{f}_{i,j}$ to that factor:

$$f_{(i,j)}(\theta) \approx \tilde{f}_{(i,j)}(\theta) \equiv \prod_{k} m_{(i,j)\to\theta_k}(\theta_k). \tag{4.7}$$

Each message can be conceptualized as the contribution of a single factor to the posterior of a single variable. If a variable is not involved with a given factor (e.g. the sender popularity $\alpha_j$ when considering the factor $i\to j$ with sender $i$ and receiver $j$), the message is uniform and provides no contribution to the posterior. Under power expectation propagation, inference proceeds by iteratively selecting a single factor $(i,j)$ and updating the messages from $(i,j)$ to the relevant variables ($\mu$, $\alpha_i$, $\beta_j$, $u_i$, and $v_j$) in order to minimize the local $\boldsymbol{\alpha}$-divergence of factor $(i,j)$, i.e. the divergence between $f_{(i,j)}\prod_{(i',j')\neq(i,j)}\tilde{f}_{(i',j')}$ and $\tilde{f}_{(i,j)}\prod_{(i',j')\neq(i,j)}\tilde{f}_{(i',j')}$. This local divergence approximates the minimization of the global $\boldsymbol{\alpha}$-divergence between the posterior $p$ and approximation $q$ under the assumption that the other factors $f_{(i',j')}$ are well-approximated by $\tilde{f}_{(i',j')}$.

As our approximation $q$ is a product of Gaussian densities, we take the messages to be unnormalized Gaussian densities, noting that these densities are closed under multiplication and we can implicitly rescale $q(\theta_k)$ to be a (normalized) Gaussian density after every iteration.

In order to minimize local $\boldsymbol{\alpha}$-divergence, the update step for variable $\theta_k$ from factor $(i,j)$

is given by

$$q'(\theta_k) = \text{proj}\left[q(\theta_k)\, m^{-\alpha}_{(i,j)\to\theta_k}(\theta_k) \int_{\theta\setminus\theta_k} f^{\alpha}_{(i,j)}(\theta) \prod_{\theta\setminus\theta_k} q(\theta)\, m^{-\alpha}_{(i,j)\to\theta}(\theta)\, d\theta\right], \tag{4.8}$$

$$q(\theta_k)^{\text{new}} = q(\theta_k)^{\epsilon}\, q'(\theta_k)^{1-\epsilon}, \tag{4.9}$$

$$m_{(i,j)\to\theta_k}(\theta_k)^{\text{new}} = \frac{q(\theta_k)^{\text{new}}\, m_{(i,j)\to\theta_k}(\theta_k)}{q(\theta_k)}. \tag{4.10}$$

where $\text{proj}[p] = \text{argmin}_q \text{KL}(p||q)$ denotes KL projection to the family of Gaussian densities (matching the mean and variance of $p$) and $\epsilon$ is a damping factor to aid with the convergence of the algorithm. Define $g(\theta_l) \equiv q(\theta_l)\, m^{-\alpha}_{(i,j)\to\theta_l}(\theta_l)$ and note $g(\theta_l)$ has the form of a Gaussian density, which we can take to be normalized. Then equation (4.8) can be written as

$$q'(\theta_k) = \text{proj}\left[g(\theta_k)\, E_{g,\theta\setminus\theta_k}\left[f^{\alpha}_{(i,j)}(\theta)\right]\right]. \tag{4.11}$$

One particular strength of the Power EP approach is the ability to choose $\boldsymbol{\alpha}$ such that evaluating the above expectations is tractable. For the logistic likelihood, the choice $\alpha = -1$, is particularly compelling[3]:

$$E_{g,\theta\setminus\theta_k}\left[f^{-1}_{(i,j)}(\theta)\right] = E_{g,\theta\setminus\theta_k}\left[1 + \exp\left(-y_{ij}\left(\mu + \alpha_i + \beta_j + u_i^T v_j\right)\right)\right] \tag{4.12}$$

$$= 1 + E_{g,\theta\setminus\theta_k}\left[\exp(-y_{ij}\mu)\exp(-y_{ij}\alpha_i)\exp(-y_{ij}\beta_j)\exp\left(-y_{ij}u_i^T v_j\right)\right] \tag{4.13}$$

where the final expectation factors over each term. There are three sets of expectations to evaluate: $E_{g,\mu}[\exp(-y_{ij}\mu)]$ (and equivalent expressions for the other univariate parameters), $E_{g,v_j}\left[\exp\left(-y_{ij}u_i^T v_j\right)\right]$, and $E_{g,u_i,v_j}\left[\exp\left(-y_{ij}u_i^T v_j\right)\right]$. The first two can be evaluated directly from the moment generating functions for the univariate and multivariate Gaussian distri-

[3]The choice of $\boldsymbol{\alpha}$ also affects the shape of the approximation $q$ relative to $p$. Minka (Minka, 2005) notes the choice of $\boldsymbol{\alpha} = -1$ puts greater emphasis on concentrating the mass of $q$ inside higher density areas of the $p$ (as opposed to "covering" the posterior) and can lead $q$ to understate the variability in the posterior.

bution, and the last can be evaluated using the independence between the distributions over $u_i$ and $v_j$ with complete the square techniques. Using $\mu_{\theta_k}$ and $\sigma_{\theta_k}$ or $\Sigma_{\theta_k}$ when appropriate to denote the mean and variance of $g(\theta_k)$:

$$E_{g,\mu}\left[\exp\left(-y_{ij}\mu\right)\right] = \exp\left(-y_{ij}\mu_\mu + \frac{1}{2}\sigma_\mu^2\right), \tag{4.14}$$

$$E_{g,v_j}\left[\exp\left(-y_{ij}u_i^T v_j\right)\right] = \exp\left(-y_{ij}\mu_{v_j}^T u_i + \frac{1}{2}u_i^T \Sigma_{v_j} u_i\right), \tag{4.15}$$

$$E_{g,u_i,v_j}\left[\exp\left(-y_{ij}u_i^T v_j\right)\right] = E_{g,u_i}\left[\exp\left(-y_{ij}\mu_{v_j}^T u_i + \frac{1}{2}u_i^T \Sigma_{v_j} u_i\right)\right] \tag{4.16}$$

$$= \det\left(\Sigma_{v_j}^{-1} - \Sigma_{u_i}\right)^{-\frac{1}{2}} \det\left(\Sigma_{v_j}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mu_{v_j}^T \Sigma_{v_j}^{-1}\mu_{v_j}\right) \tag{4.17}$$

$$\times \exp\left(\frac{1}{2}\left(\mu_{u_i} + \Sigma_{v_j}^{-1}\mu_{v_j}\right)^T \left(\Sigma_{v_j}^{-1} - \Sigma_{u_i}\right)^{-1} \left(\mu_{u_i} + \Sigma_{v_j}^{-1}\mu_{v_j}\right)\right). \tag{4.18}$$

The complete the square used in the last equation relies on $\Sigma_{v_j}^{-1} - \Sigma_{u_i}$ being a positive definite matrix in order for the resulting density to be a multivariate normal distribution.

We focus on the update steps for $\alpha_i$ and $u_i$, noting the symmetry in equation (4.13) with respect to $\mu$, $\alpha_i$, and $\beta_j$, and similarly for $u_i$ and $v_j$, implies the corresponding update steps can be obtained to swapping the positions of the relevant variables. The updates take the form

$$q'\left(\alpha_i\right) = \text{proj}\left[g\left(\alpha_i\right)\left(1 + c_1 \exp\left(-y_{ij}\alpha_i\right)\right)\right], \tag{4.19}$$

$$q'\left(u_i\right) = \text{proj}\left[g\left(\alpha_i\right)\left(1 + c_2 \exp\left(-y_{ij}\mu_{v_j}^T u_i + \frac{1}{2}u_i^T \Sigma_{v_j} u_i\right)\right)\right], \tag{4.20}$$

with

$$c_1 = E_\mu\left[\exp\left(-y_{ij}\mu\right)\right] E_{\beta_j}\left[\exp\left(-y_{ij}\beta_j\right)\right] E_{u_i,v_j}\left[\exp\left(-y_{ij}u_i^T v_j\right)\right],$$

$$c_2 = E_\mu\left[\exp\left(-y_{ij}\mu\right)\right] E_{\alpha_i}\left[\exp\left(-y_{ij}\alpha_i\right)\right] E_{\beta_j}\left[\exp\left(-y_{ij}\beta_j\right)\right].$$

These densities can be represented as a linear combination of two Gaussian densities:

$$q'(\alpha_i) = \text{proj}\left[N\left(\alpha_i; \mu_{\alpha_i}, \sigma^2_{\alpha_i}\right) + cN\left(\alpha_i; \mu_{\alpha_i} - y_{ij}\sigma^2_{\alpha_i}, \sigma^2_{\alpha_i}\right)\right], \tag{4.21}$$

$$q'(u_i) = \text{proj}\left[N(u_i; \mu_g, \Sigma_g) + cN\left(u_i; \left(\Sigma^{-1}_{u_i} - \Sigma_{v_j}\right)^{-1}\left(\Sigma^{-1}_{u_i}\mu_{u_i} - y_{ij}\mu_{v_j}\right), \left(\Sigma^{-1}_{u_i} - \Sigma_{v_j}\right)^{-1}\right)\right], \tag{4.22}$$

where

$$c = E_\mu\left[\exp\left(-y_{ij}\mu\right)\right] E_{\alpha_i}\left[\exp\left(-y_{ij}\alpha_i\right)\right] E_{\beta_j}\left[\exp\left(-y_{ij}\beta_j\right)\right] E_{u_i,v_j}\left[\exp\left(-y_{ij}u_i^T v_j\right)\right] \tag{4.23}$$

and first and second moments can be calculated from each expression (after normalizing by $\frac{1}{1+c}$) to derive the corresponding Gaussian parameters.

To recap, our message passing algorithm will proceed as follows:

1. Initialize all messages $m_{(i,j)\to\theta_k}$
2. Repeat until convergence of all messages:
   (a) Choose factor $(i, j)$
   (b) Update approximation to posterior $q(\theta)$ via equations (4.8) and (4.9). We find the choice of $\epsilon = 2$ promising in simulations.
   (c) Update messages from this factor $m_{(i,j)\to\theta_k}$ via equation (4.10)

### *4.2.3 Case-Control Approximation*

The update step for each factor has $O(1)$ computational cost, but each iteration over the entire network has $O(N^2)$ computational cost and can be prohibitively expensive in large networks. In addition, tracking the messages for each factor also has $O(N^2)$ storage complexity. Drawing inspiration from (Raftery et al., 2012), we wish to take advantage of the idea that large networks tend to be sparse and iterating over the entire network can be computationally inefficient due to the extreme class imbalance. The influence contained in

each non-edge may be relatively small towards informing the overall model compared to the influence of edges, which are many fewer in number.

We propose iterating over the set of factors with $y_{i,j} = 1$ and a random sample of factors with $y_{i,j} = 0$[4]. Supposing the number of observed edges $E = |\{i, j\} : y_{i,j} = 1| << N^2$ and a random sample of non-edges of size $O(E)$ is drawn, each iteration over the network would cost $O(E)$ rather than $O(N^2)$. Algorithmically, we treat this sample of factors as if it is the full set of data available. To understand the effect of this choice on the means of our parameter estimates, consider exponentiating both sides of equation (4.2):

$$\frac{p_{i,j}}{1 - p_{i,j}} = \exp(\mu + \alpha_i + \beta_j + u_i^T v_j). \tag{4.24}$$

Intuitively, sampling a random proportion $q$ of the non-edges inflates the odds-ratio on the LHS by a factor of $q^{-1}$. On the RHS, $\mu$ would shift upwards by $-\log(q)$ but the other parameters would be unaffected. This suggests a simple post-hoc mean correction would suffice to return the parameters to their original scale, although it should be noted the posterior variance of the latent variables should be larger than if the full dataset were used. We provide some evidence for the efficacy of this case-control approximation in Section 4.3.

### *4.2.4 Temporal Dynamics*

Next, we introduce discrete time dynamics to the bilinear mixed-effects model by allowing each of the latent variables to evolve via a Markov chain. Let $\theta_{t,k}$ denote the $k$th parameter at time $t$, with the posterior of $\theta_{t,k}$ given (approximated) by

$$\theta_{t,k} | \mathbf{Y}_{1:t} \sim N(\widehat{\mu}_{\theta_{t,k}}, \widehat{\Sigma}_{\theta_{t,k}}). \tag{4.25}$$

[4]In practice, drawing a single random sample is preferable to drawing a new sample each iteration through the data due to the reduced time observed for the convergence of the algorithm. In a temporal setting, a new sample can be drawn at each time step.

Supposing $\theta_{t,k}$ evolves via the Gaussian random walk

$$\theta_{t+1,k} \sim \theta_{t,k} + N(0, W_{t+1,k}), \tag{4.26}$$

the prior for $\theta_{t+1,k}$ would be given by

$$\theta_{t+1,k}|\mathbf{Y}_{1:t} \sim N(\widehat{\mu}_{\theta_{t,k}}, \widehat{\Sigma}_{\theta_{t,k}} + W_{t+1,k}). \tag{4.27}$$

We adopt the adaptive tuning procedure described in (McCormick et al., 2012) to determine the amount of additional variation to introduce at each time point. We parametrize this amount of variation via a "forgetting" multiplier $\tau_{t+1,k} \geq 1$ to avoid specifying $d \times d$ random walk matrices for $u_i$ and $v_j$:

$$\theta_{t+1,k}|\mathbf{Y}_{1:t} \sim N(\widehat{\mu}_{\theta_{t,k}}, \tau_{t+1,k}\widehat{\Sigma}_{\theta_{t,k}}). \tag{4.28}$$

We choose these multipliers $\boldsymbol{\tau}_{t+1}$ based on the average predictive likelihood:

$$\boldsymbol{\tau_{t+1}} = \underset{\boldsymbol{\tau}_{t+1}}{\text{argmax}} \frac{1}{N^2} \sum_{i,j} \int_{\theta_{t+1}} p(y_{i,j}|\theta_{t+1}, \mathbf{Y}_{1:t}) p(\theta_{t+1}|\mathbf{Y}_{1:t}) d\theta_{t+1}. \tag{4.29}$$

Evaluating the integral above cannot be done in closed form, and we use a series of two approximations to estimate it. First, note the likelihood term primarily involves the sigmoid of the logistic mean function:

$$p(1|\theta_{t+1}, \mathbf{Y}_{1:t}) = p_{t+1} = \text{expit}(\mu_{t+1} + \alpha_{t+1,i} + \beta_{t+1,j} + u_{t+1,i}^T v_{t+1,j}, \tag{4.30}$$

$$p(0|\theta_{t+1}, \mathbf{Y}_{1:t}) = 1 - p_{t+1}. \tag{4.31}$$

Recall the prior over each latent variable is an independent Gaussian. We approximate

$u_{t+1,i}^T v_{t+1,j}$ with a single Gaussian term (via their first two moments) in order to model the entire mean function itself with a single Gaussian. Denoting the mean function with $\psi$, we use the following approximation for convoluting a sigmoid and a Gaussian could be used (see (Bishop, 2006)):

$$\int \text{expit}(\psi) N(\psi|\mu_\psi, \sigma_\psi^2) = \text{expit}((1 + \pi\sigma_\psi^2/8)^{-1/2}\mu_\psi). \tag{4.32}$$

Lastly, in order to reduce the computational cost involved in maximizing (4.29), we allow for a single forgetting multiplier for $\mu_{t+1}$, a single multiplier for the popularity terms $\alpha_{t+1,i}$ and $\beta_{t+1,j}$, and a single multiplier for the latent space terms $u_{t+1,i}$ and $v_{t+1,j}$, and only evaluate (4.29) over a coarse grid of values. McCormick et al. (2012) argue searching over a coarse grid[5] leads to comparable results to directly maximizing (4.29) when running the algorithm over sufficiently many time periods, as periods with unnecessary inflation in prior variance can be balanced against periods with more restrictive inflation. In Sections 4.3 and 4.4, we take $\tau \in \{1, 1.01, 1.1, 2\}$. This choice of values allows for no change in a paramter ($\tau = 0$), as well as forgetting multipliers corresponding to multiple scales of variance inflation.

### *4.2.5 Anomaly Detection*

At an edge level, anomaly detection proceeds by scoring edges based on their probabilities for observing activity, under the assumption that any past activity is non-anomalous and thus is a good representation of normal behavior. We score the dyad $i \to j$ at time $t$ via its predictive likelihood:

$$\hat{p}_{i,j,t+1} \equiv \int_{\theta_{t+1}} p(y_{i,j}|\theta_{t+1}, \mathbf{Y}_{1:t}) p(\theta_{t+1}|\mathbf{Y}_{1:t}) d\theta_{t+1}, \tag{4.33}$$

[5]Furthermore, the authors found their results were robust to the choice of grid values.

which can be evaluated via the approximations described in the previous section. Dyads with activity but low predictive scores as well as dyads without activity but high predictive scores would then be flagged as anomalous.

For many settings, we may be interested in detecting anomalies at a non-edge level. For example, in computer networks such as the LANL network described in Section 4.1, security experts are interested in identifying anomalous subgraphs which may potentially represent intruder attacks. (Neil et al., 2013a) mentions activity in the shape of $k$-stars and $k$-paths as common behavior for intrusions. Note dyads in these subgraphs would consist solely of edges with observed activity, so lower values of $\hat{p}_{i,j,t+1}$ would be characterized as more anomalous. We can compute scores for these subgraphs from our edge level scores given a conditional independence assumption, by multiplying the scores of the corresponding edges, or equivalently, summing the log scores. For example, for the 3-path shown in figure 4.1, the

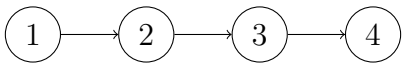


Figure 4.1: A directed 3-path.

score would be given by $\hat{p}_{1,2,t+1}\hat{p}_{2,3,t+1}\hat{p}_{3,4,t+1}$.

In this chapter, we choose to separately consider potentially anomalous behavior for each time period, although combining scores across time may be promising, particularly in settings with fine temporal resolution where attacks may span multiple periods. A fully online detection procedure would proceeds at each time step as follows:

1. Observe network behavior $\mathbf{Y}_t$
2. Tune forgetting multipliers (4.29)
3. Flag and assess potential anomalous subgraphs
4. Remove anomalous activity from $\mathbf{Y}_t$

5. Estimate model parameters $\Theta_t$

### *4.3 Simulation Study*

To provide an idea of how well our proposed algorithm can estimate a dynamic bilinear mixed-effects model, we simulate a network following equations (4.1) and (4.2) and with time dynamics following independent Gaussian random walks (see (4.26)). Specifically, we generate a network of size $N = 500$ from a bilinear mixed-effects model with latent dimension 2 with the following priors:

$$\mu_1 \sim N(-6.5, 0.1), \tag{4.34}$$

$$\alpha_{1,i} \sim N(0, 1), \tag{4.35}$$

$$\beta_{1,j} \sim N(0, 1), \tag{4.36}$$

$$u_{1,i} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}0.75 & 0.15\\0.15 & 0.75\end{bmatrix}\right), \tag{4.37}$$

$$v_{1,j} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}0.75 & 0.15\\0.15 & 0.75\end{bmatrix}\right). \tag{4.38}$$

We evolve the network 99 times for a total of $T = 100$ periods, where at every time point each parameter follows a Gaussian random walk with (co)variance equal to 0.001 times its prior variance. These prior and random walk values were chosen to create a network that would be roughly similar to the LANL computer network, that is, characterized by high sparsity, strong heterogeneity between nodes, low temporal dynamics, and strong dependence between time periods. The generated network averages about 2,000 directed connections per time period, or 4 per node, which is slightly less than what we observe for the LANL network.

We compare results from two runs of the Power EP algorithm described in Section 4.2.2, one that iterates over all 250,000 potential dyads in the network and another that implements the case-control modification described in Section 4.2.3. For the latter, we sample 2.5% of the non-edges at every time point for consideration, thus iterating over about 8,200 dyads per time point. This results in about a 93.5% reduction in computation time and a 97% reduction in storage complexity.

We compare mean estimates from each run against the generated values in terms of log-likelihood, area under the receiver operating characteristic curve (ROC AUC), and the correlation between the actual and estimated edges probabilities on the logit scale. This correlation is also calculated restricted to dyads not observed in any of the 100 time periods (which is satisfied by 72.6% of all dyads). In Figure 4.2, we compare the model fit of the mean estimates from the Power EP algorithm with and without the case-control approximation to the model fit of the true parameter values. In these plots, the log-likelihood and ROC AUC under the generating model provide a soft bound on model performance, as no other set of parameter estimates should be systematically outperform them over a prolonged period. The estimates from our variational approach perform very similarly to the generating model, particularly after time period 40, suggesting the approximations used in variational method have, at most, minor effects on the mean posterior estimates.

In Figure 4.3, we plot the correlation between the actual edge probabilities and their estimated counterparts on the logit scale. The performance of the algorithms ramp up over time, as each binary network $\mathbf{Y}$ provides limited information about the underlying latent variables which must be aggregated, and is largely stabilized by time 40. The case-control modified algorithm, which iterates over a much smaller subset of the network at each time point, does perform worse than the algorithm over the full network, but these differences are largest in the earlier time periods.

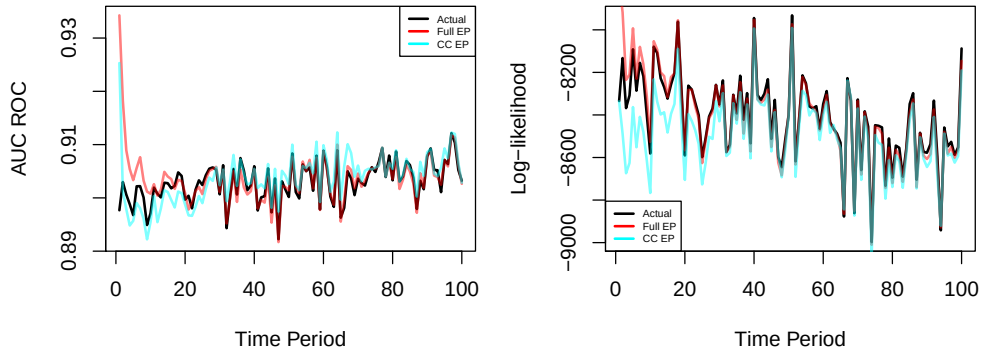


Figure 4.2: Log-likelihood of the observed network $\mathbf{Y}_t$ and ROC AUC of the edge probabilities $\hat{p}_{i,j,t}$. Calculated for three sets of edge probabilities: the actual edge probabilities (black), the edge probabilities estimated via the Power EP algorithm on the full data (red), and the edge probabilities estimated via the Power EP algorithm with the case-control approximation (cyan).

Restricting ourselves to results from three time points, we plot the distributions of the edge probabilities (again on the logit scale) against their actual counterparts in Figure 4.4, and find minor systematic differences between the distributions. Note both sets of estimated probabilities do struggle a bit (overestimating) modeling the extreme left tail of probabilities, although these differences are exacerbated due to the logit scale (e.g. expit$(-14) =$ 8.3e-07 and expit$(-16) =$ 1.1e-07) and may be hard to capture given the time frame of the simulation in comparison to the probability size.

Lastly, we find the increase in posterior variance of our parameters when adopting the case-control modification to be largely acceptable. Even though we only consider about 3% of the edges in any given time period, this subset of the network appears to capture most of the information for estimating the model parameters.

## *4.4 LANL Netflow Event Data*

We demonstrate the potential for anomaly detection with Netflow communications data on the LANL enterpise network (Turcotte et al., 2018). Event records correspond to directed

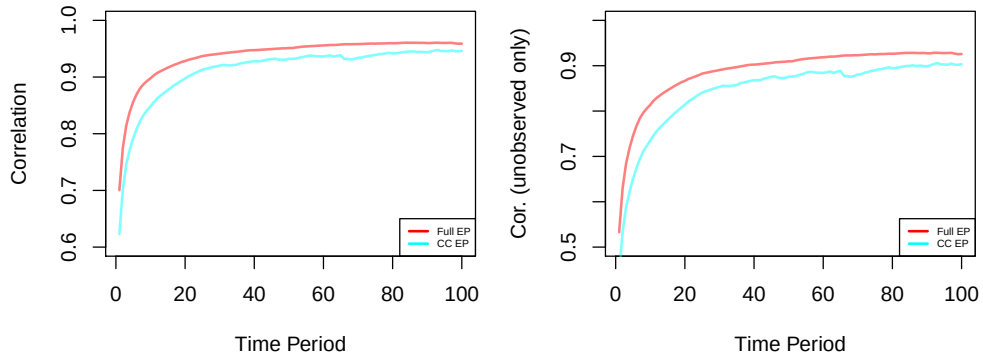


Figure 4.3: Correlation between the actual edge probabilities and estimated edge probabilities, both on the logit scale. Results with probabilities from the full data are in red, and results with probabilities from the case-control are in cyan.

| Time Period | $\mu$ | $\alpha_i$ | $\beta_j$ | $u_{i,1}$ | $v_{j,2}$ |
|---|---|---|---|---|---|
| T = 10 | 6.76 | 1.92 | 1.97 | 2.11 | 2.42 |
| T = 25 | 1.92 | 2.47 | 2.55 | 2.10 | 2.32 |
| T = 100 | 2.24 | 2.63 | 2.73 | 1.54 | 1.78 |

Table 4.1: Multiplier in posterior variance when using the case-control modification to the Power EP algorithm. For node-specific parameters, the multiplier in variance is calculated for each node and averaged.

communication between two network devices and span a total of 89 days. We restrict to the sub-network of the $N = 27,436$ computers with some record of outgoing communications over the 89 days[6], and aggregate the event records into four-hour intervals, yielding a total of $T = 532$ time periods. We focus on modeling the presence of any directed network activity between each dyad within each four-hour interval. The resulting network averages about 150,000 directed edges (or 5.5 outgoing edges for each computer) at each time interval, and there is substantial variation in activity levels based on time-of-day and day-of-week.

The LANL data contains a red team attack in the form of a network scanning attack from "Computer A" that begins on day 57, and we are interested in the ability for our model to recognize this activity as anomalous. Following (Neil et al., 2013a), we detect

[6]These computers comprise the set of network devices which may be the source of malicious behavior.

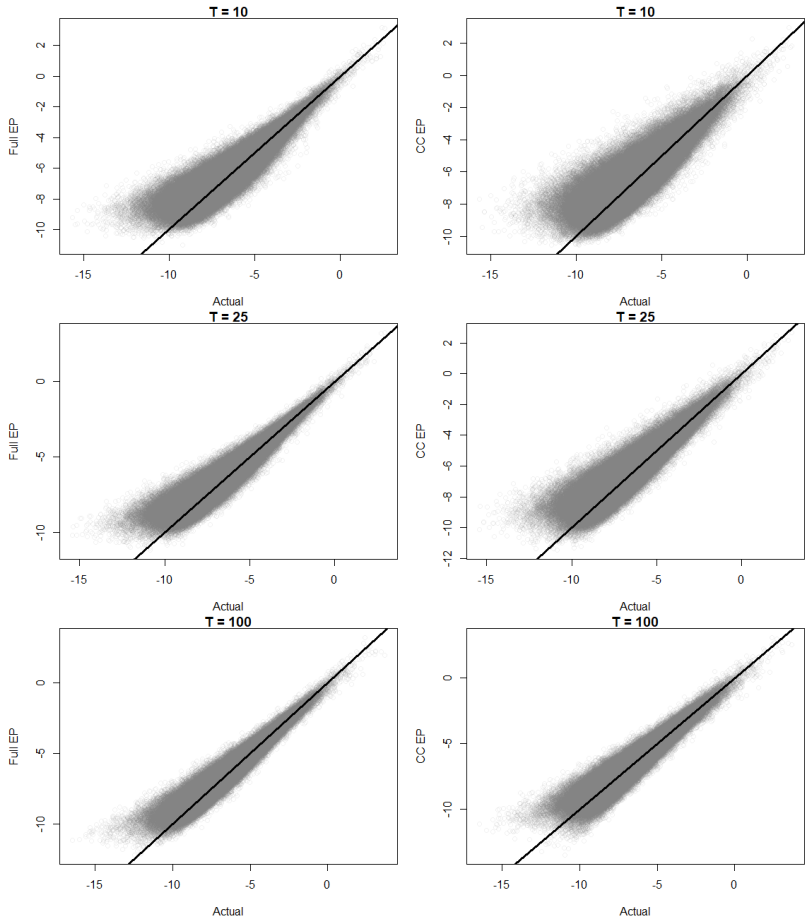


Figure 4.4: Actual versus estimated edge probabilities on the logit scale. Left column compare the actual edge probabilities to estimates from the full Power EP, while the right column represents the edge probabilities estimated via the case-control Power EP.

potentially anomalous subgraphs of the three shapes presented in Figure 4.5, corresponding to common intrusion patterns. While malicious attacks may involve more nodes and activity, detecting a single subgraph involved in the attack may suffice to identify the entire attack upon further (manual) examination. Note the detection procedure described in this section deviates from the typical online setting since details of the red team attack were only obtained after model estimation. Flagging potential anomalies only occurs after model estimation, and the estimated probabilities used in this process assume all preceding network activity was

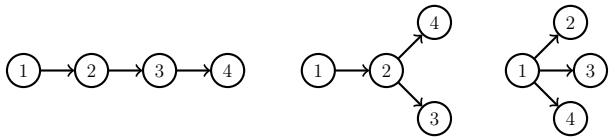


Figure 4.5: A 3-path (left), a 3-star (right), and a "fork" (center) representing a combination of the two.

non-anomalous.

We estimate the bilinear mixed-effects model with latent dimension $d = 2$ using the Power EP approach described in Section 4.2.2 with the case-control approximation of Section 4.2.3, taking a sample of the non-edges of average size 500,000 (corresponding to a case-control rate of 3.3 or sampling proportion $q \approx 0.066\%$). We slightly modify the bilinear mixed-effects model to incorporate time-of-day and day-of-week effects in the form of mean shifts, with individual terms for each time-of-day and day-of-week pair calculated directly from the mean activity levels over the 89 days[7]. We choose to separately model these terms from the overall popularity term $\mu_t$ in order to prevent the dynamics of this parameter to be governed by periodicity effects rather than random walk behavior. Note that part of the periodicity effects may be due to recurrent, automated tasks (e.g. weekly at a certain time of day), so allowing for more complicated periodicity effects or removing these activities before estimation (if they are labeled or can be *a priori* identified) would likely improve model fit.

Before turning to anomaly detection, we assess how well the bilinear mixed-effects model and the popularity model omitting the latent interaction terms are able to predict LANL network activity. Figure 4.6 plots the area under the receiver operating characteristic curve (AUC ROC) of both models calculated using probabilities from the predictive likelihood (4.33), which are primarily dependent on estimated parameters from the previous time pe-

[7]A slightly more sophisticated approach would be to include them as additional parameters in the model to estimate. This would allow these effects to naturally change over time, although there is little evidence for any such changes in the observed data.

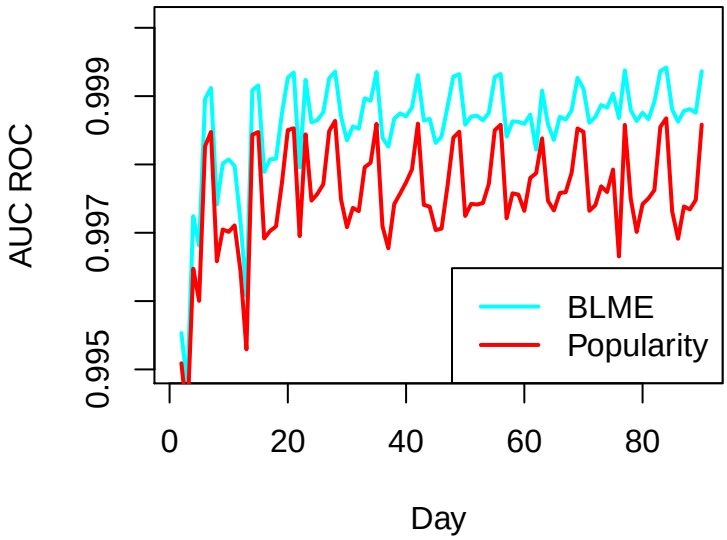


Figure 4.6: Averaged AUC ROC for popularity model (red) and bilinear mixed-effects model (cyan) over days. AUC ROC is calculated for each 4-hour interval using probabilities derived from the predictive likelihood and results are averaged across each day.

riod. Both models perform quite well, with AUC $> 0.995$, suggesting the network communications data is inherently very structured and predictable in nature. Model performance exhibits both time-of-day and day-of-week periodicity, suggesting a more complex approach to modeling periodicity is likely to improve performance, albeit performance is consistently high despite these effects. The latent interaction terms in the bilinear mixed-effects model do seem to substantially improve performance, with these improvements primarily driven by higher probabilities for active dyads with generally low overall levels of popularity in the network.

We can compute anomaly scores for subgraphs of the types shown in Figure 4.5 by taking the sum of the log probabilities as described in Section 4.2.5 . Despite the relative sparsity of the network, the number of subgraphs to consider at each time frame remains quite large. To reduce the number of subgraphs in consideration further, we only examine subgraphs consisting of edges with log probability score of -10 or lower and remove some "overlapping"

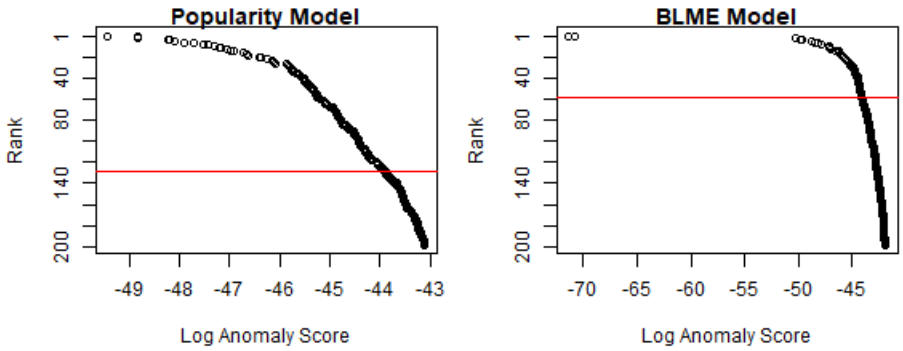


Figure 4.7: Anomaly scores for the 200 lowest scoring subgraphs observed over the 89 day period. The red line corresponds to the rank of a subgraph containing part of a red team attack on day 57.

subgraphs [8]. In Figure 4.7, we plot the 200 most anomalous subgraphs under each model. Both sets of subgraphs contain a single 3-star (highlighted in red) involving the network scanning attack from Computer A on the first day of the red team attack. The rank of the Computer A 3-star is twice as high under the bilinear mixed-effects model, where this anomaly would be detectable given an average alarm rate of one subgraph per day. The difference in rank can be mainly attributed to low scores on recurrent activity between low-popularity computers under the popularity model, which is not flexible enough to model such activity. This leads to lower scores for certain non-anomalous subgraphs, obfuscating the actual attack. Note our detection procedure is largely unable to identify other attacks from Computer A in the following days. Once anomalous behavior like the activity on day 57 is incorporated into the model of normal activity, subsequent attacks appear to be normal behavior.

[8]Specifically, we remove 3-paths with the same middle edge "2" $\rightarrow$ "3", forks with the same "2" node, and 3-stars with the same "1" node. Ideally, detecting one anomaly from multiple overlapping subgraphs would suffice for finding the entire attack.

## *4.5 Discussion*

In this chapter, we present an variational approach for estimating the bilinear mixed-effects model. We adapt our approach to a dynamic, large network setting via a case-control approximation and an autotuning procedure to mimic Gaussian random walks on the model parameters. We demonstrated the efficacy of our algorithm via a simulation study on $N = 500$ nodes, estimated the mixed-effects model on the LANL netflow communications network involving over 25,000 computers for a period of 89 days, and detected a red team attack on the same network while only requiring half the detection rate of the popularity model.

A natural extension for the bilinear mixed-effects model considered would be to allow for node- or edge-level covariates in the specification of the mean function. Relational event data is often provided with additional details which may be useful in conjunction with network-based predictors for predicting activity. Along a similar line of reasoning, edge-level covariate data may distinguish between multiple types of network activity which we may wish to model jointly. In particular, the LANL netflow data includes sender port information, and utilizing this information would help distinguish between typical activity on a commonly used port and unusual activity on an rarely used port. Lastly, adapting the algorithm to handle different outcome measures may allow for a more faithful representation of the observed event data. For example, following Chapter 3 in acknowledging the data's continuous-time nature, modeling the communication activity using Poisson processes with time dependent intensities could exploit detailed information about the timing of expected activity.

# Chapter 5

# CONCLUSION AND FUTURE DIRECTIONS

In this dissertation, we highlighted several challenges associated with utilizing network data from both survey sources and relational event logs. We contested that survey data is prone to mismeasuring the true relational network of interest due to both human error and survey design choices. In Chapter 2, we demonstrated how this mismeasurement might manifest when estimating treatment effects of experiments in networks. We then turned our attention to relational event data, which are often able to accurately and inexpensively measure activity between actors in a network. However, additional modeling is required in order to construct a relational network from this data. In Chapters 3 and 4, we argued that existing network models were inadequately able to handle certain properties characteristic of relational event data, most notably its continuous-time nature and large scale.

Throughout this work, our fundamental approach to addressing these challenges was to explicitly model the relationship between the observed data and the underlying relational network of interest. In Chapter 2, we incorporated the uncertainty due to the potential mismeasurement of the latent network into the estimation of treatment effects, while in Chapters 3 and 4 we modeled the relational event data as stochastic manifestations of underlying social structure. However, as our network of interest is inherently unobserved, we were forced to make significant modeling assumptions in order to define a relationship between the observed and hypothesized latent network. In Chapters 3 and 4, we related the two networks with highly specific parametric forms, while in Chapter 2 we also forwent the design-based estimator in favor of a model-based estimator requiring parametric forms for the outcomes.

The work presented in this thesis can be extended in a few different directions. While we were able to separately address two concerns regarding modeling relational event data in Chapters 3 and 4, the ideal approach would be to jointly address both concerns in a unified framework. An interesting starting point might be to incorporate the latent factor model into the intensity of a Poisson process model; the use of point processes most closely reflects the continuous-time nature of activity data, while the use of node-specific parameters can aid with scalability. The challenge would then be an issue of modeling the evolution of the latent parameters in order to simultaneously permit estimation and appropriate temporal dynamics. For the mixture model proposed in Chapter 2, a natural extension might seek to consider more complicated parametric forms for potential outcomes under each exposure condition. Specifically, allowing for each subject's outcomes to depend on other observed covariates could address potential concerns of random imbalance in the treatment conditions.

# BIBLIOGRAPHY

Acemoglu, D., Ozdaglar, A., and Yildiz, E. (2011). Diffusion of innovations in social networks. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 2329–2334.

Adhvaryu, A., Kala, N., and Nyshadham, A. (2019). Management and shocks to worker productivity. Technical report, National Bureau of Economic Research.

Angelucci, M., De Giorgi, G., Rangel, M. A., and Rasul, I. (2010). Family networks and school enrolment: Evidence from a randomized social experiment. *Journal of Public Economics*, 94(3-4):197–221.

Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947.

Bandiera, O., Barankay, I., and Rasul, I. (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica*, 77(4):1047–1094.

Banerjee, A., Breza, E., Chandrasekhar, A. G., and Mobius, M. (2019). Naive learning with uninformed agents. Technical report, National Bureau of Economic Research.

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

Basford, K. E., Greenway, D. R., McLachlan, G. J., and Peel, D. (1997). Standard errors of fitted component means of normal mixtures. *Computational Statistics*, 12(1):1–17.

Beaman, L. A., BenYishay, A., Magruder, J., and Mobarak, A. M. (2018). Can network theory-based targeting increase technology adoption? Technical report, National Bureau of Economic Research.

Bell, D. C., Belli-McQueen, B., and Haider, A. (2007). Partner naming and forgetting: Recall of network members. *Social Networks*, 29(2):279–299.

BenYishay, A. and Mobarak, A. M. (2018). Social learning and incentives for experimentation and communication. *Review of Economic Studies*, 86(3):976–1009.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.

Blattman, C., Green, D., Ortega, D., and Tobón, S. (2017). Place-based interventions at scale: The direct and spillover effects of policing and city services on crime. Technical report, National Bureau of Economic Research.

Bloom, N., Liang, J., Roberts, J., and Ying, Z. J. (2014). Does working from home work? Evidence from a Chinese experiment. *The Quarterly Journal of Economics*, 130(1):165–218.

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. *Information Technology and Development*, 18(2):107–125.

Blundell, C., Heller, K. A., and Beck, J. M. (2012). Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems 25*.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55.

Bursztyn, L., Ederer, F., Ferman, B., and Yuchtman, N. (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, 82(4):1273–1301.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38(1):155–200.

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.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Li, P., and Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

Centola, D. and Macy, M. (2007). Complex contagions and the weakenss of long ties. *American Journal of Sociology*, 113(3):702–734.

Chandrasekhar, A. G. and Lewis, R. (2011). Econometrics of Sampled Networks.

Conley, T. G. and Udry, C. R. (2010). Learning about a new technology: Pineapple in Ghana. *American Economic Review*, 100(1):35–69.

Conlon, C. T. and Mortimer, J. H. (2013). Efficiency and foreclosure effects of vertical rebates: Empirical evidence. Technical report, National Bureau of Economic Research.

Crook, J. R. and Shields, W. M. (1987). Non-parental nest attendance in the barn swallow (Hirundo rustica): helping or harassment? *Animal Behaviour*, 35(4):991–1001.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38.

Dong, W., Pentland, A. S., and Heller, K. A. (2012). Graph-coupled HMMs for modeling the spread of infection. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 227–236.

DuBois, C. and Smyth, P. (2010). Modeling relational events via latent classes. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 803–812.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, 101(5):1739–1774.

Durante, D. and Dunson, D. B. (2016). Locally adaptive dynamic networks. *The Annals of Applied Statistics*, 10(4):2203–2232.

Eagle, N., Pentland, A. S., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15274–15278.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer, New York.

Godlonton, S. and Thornton, R. (2012). Peer effects in learning HIV results. *Journal of Development Economics*, 97(1):118–129.

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

Griffin, M., Gile, K. J., Fredricksen-Goldsen, K. I., Handcock, M. S., and Erosheva, E. A. (2018). A simulation-based framework for assessing the feasibility of respondent-driven sampling for estimating characteristics in populations of lesbian, gay and bisexual older adults. *The Annals of Applied Statistics*, 12(4):2252–2278.

Griffith, A. (2017). Random assignment with non-random peers: A structural approach to counterfactual treatment assessment. Technical report.

Grün, B. and Leisch, F. (2007). Finite mixtures of generalized linear regression models. *Recent Advances in Linear Models and Related Areas*, pages 205–230.

Handcock, M. S. and Gile, K. J. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25.

Hardy, M. and McCasland, J. (2019). It takes two: Experimental evidence on the determinants of technology diffusion. From Duplicate 1 (It Takes Two : Experimental Evidence on the Determinants of Technology Diffusion - Hardy, Morgan; McCasland, Jamie) NULL.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Heath, R. (2018). Why do firms hire using referrals? Evidence from Bangladeshi garment factories. *Journal of Political Economy*, 126(4):1691–1746.

Hennig, C. (2000). Identifiablity of models for clusterwise linear regression.

Hoff, P. D. (2005). Bilinear mixed effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Howison, J., Wiggins, A., and Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, 12(12):767–797.

Killworth, P. D. and Bernard, H. R. (1976). Informant accuracy in social network data. *Human Organization*, 35(3):269–286.

Krivitsky, P. N. and Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B*, 76(1):29–46.

Levin, I. I., Zonana, D. M., Burt, J. M., and Safran, R. J. (2015). Performance of Encounternet tags: Field tests of miniaturized proximity loggers for use on small birds. *PLoS ONE*, 10(9).

Levin, I. I., Zonana, D. M., Fosdick, B. K., Song, S. J., Knight, R., and Safran, R. J. (2016). Stress response, gut microbial diversity and sexual signals correlate with social interactions. *Biology Letters*, 12(6).

Linderman, S. W. and Adams, R. P. (2014). Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1413–1421.

Madan, A., Cebrian, M., Moturu, S. T., Farrahi, K., and Pentland, A. S. (2012). Sensing the health state of a community. *IEEE Pervasive Computing*, 11(4):36–45.

Madan, A., Moturu, S. T., Lazer, D., and Pentland, A. S. (2010). Social sensing: Obesity, unhealthy eating and exercise in face-to-face networks. In *Wireless Health 2010*, pages 104–110.

Magruder, J. R. (2010). Intergenerational networks, unemployment, and persistent inequality in South Africa. *American Economic Journal: Applied Economics*, 2(1):62–85.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):1–23.

Marsden, P. V. (2016). Survey methods for network data. In Scott, J. and Carrington, P., editors, *The SAGE Handbook of Social Network Analysis*, pages 370–388. SAGE Publications, London.

McCormick, T. H., Raftery, A. E., Madigan, D., and Burd, R. S. (2012). Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics*, 68(1):23–30.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. Marcel Dekker, New York.

Metelli, S. and Heard, N. (2016). Model-based clustering and new edge modelling in large computer networks. In *2016 IEEE Conference on Intelligence and Security Informatics*, pages 91–96.

Miguel, E. and Kremer, M. (2004). Worms: I dentifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217.

Minka, T. P. (2005). Divergence measures and message passing. Technical report, Microsoft.

Moller, A. P., Barbosa, A., Cuervo, J. J., de Lope, F., Merino, S., and Saino, N. (1998). Sexual selection and tail streamers in the barn swallow. *Proceedings of the Royal Society B: Biological Sciences*, 265(1394):409–414.

Neil, J., Hash, C., Brugh, A., Fisk, M., and Storlie, C. B. (2013a). Scan statistics for the online detection of locally anomalous subgraphs. *Technometrics*, 55(4):403–414.

Neil, J. C., Uphoff, B. D., Storlie, C. B., and Hash, C. L. J. (2013b). Towards improved detection of attackers in computer networks: New edges, fast updating, and host agents. In *2013 6th International Symposium on Resilient Control Systems*.

Newman, M. E. J. (2018). Network structure from rich but noisy data. *Nature Physics*, 14(June):1–4.

Oster, E. and Thornton, R. (2011). Menstruation, sanitary products, and school attendance: Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 3(1):91–100.

Oster, E. and Thornton, R. (2012). Determinants of technology adoption: Peer effects in menstrual cup take-up. *Journal of the European Economic Association*, 10(6):1263–1293.

Perry, P. O. and Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B*, 75(5):821–849.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rutz, C. and Hays, G. C. (2009). New frontiers in biologging science. *Biology Letters*, 5(3).

Sacerdote, B. I. (2001). Peer effects with random assignment: Results for Darthmouth roommates. *The Quarterly Journal of Economics*, 116(2):681–704.

Sadilek, A., Kautz, H., and Silenzio, V. (2012). Modeling spread of disease from social interactions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, pages 322–329.

Safran, R. J. and McGraw, K. J. (2004). Plumage coloration, not length or symmetry of tail-streamers, is a sexually selected trait in North American barn swallows. *Behavioral Ecology*, 15(3):455–461.

Safran, R. J., Neuman, C. R., McGraw, K. J., and Lovette, I. J. (2005). Dynamic paternity allocation as a function of male plumage color in barn swallows. *Science*, 309(5744):2210–2212.

Salter-Townshend, M. and Murphy, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Computational Statistics and Data Analysis*, 57(1):661–671.

Scharf, H. R., Hooten, M. B., Fosdick, B. K., Johnson, D. S., London, J. M., and Durban, J. W. (2016). Dynamic social networks based on movement. *The Annals of Applied Statistics*, 10(4):2183–2202.

Sekara, V., Stopczynski, A., and Lehmann, S. (2016). Fundamental structures of dynamic

social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 113(36):9977–9982.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657.

Simma, A. and Jordan, M. I. (2010). Modeling events with cascades of Poisson processes. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 546–555.

Snijders, T. A. (1996). Stochastic actor-oriented models for network change. *The Journal of Mathematical Sociology*, 21(1-2):149–172.

Snijders, T. A. (2005). Models for longitudinal network data. In Carrington, P. J., Scott, J., and Wasserman, S., editors, *Models and Methods in Social Network Analysis*, pages 215–247. Cambridge University Press, New York.

Sobel, M. E. and Muthén, B. (2012). Compliance mixture modelling with a zero-effect complier class and missing data. *Biometrics*, 68(4):1037–1045.

Sulo, R., Berger-Wolf, T., and Grossman, R. (2010). Meaningful selection of temporal resolution for dynamic networks. In *Proceedings of the 8th Workshop on Mining and Learning with Graphs*, pages 127–136.

Turcotte, M. J. M., Kent, A. D., and Hash, C. (2018). Unified host and network data set. In Heard, N., Adams, N., Rubin-Delanchy, P., and Turcotte, M., editors, *Data Science for Cyber-Security*, chapter 1, pages 1–22. World Scientific, London.

Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337.

Vasilaky, K. N. and Leonard, K. L. (2018). As good as the networks they keep? Improving outcomes through weak ties in rural Uganda. *Economic Development and Cultural Change*, 66(4):755–792.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.

Wang, S.-Y. (2013). Marriage networks, nepotism, and labor market outcomes in China. *American Economic Journal: Applied Economics*, 5(3):91–112.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19.

Watts, D. J. (2007). A twenty-first century science. *Nature*, 445(7127):489.

Williams, K. M. (2016). *Economic analyses of educational achievement*. PhD thesis, University of California, Davis.