Causal Inferential Dynamic Network Analysis

by

Sebastián Martínez Bustos

A dissertation submitted to the
University of Glasgow
for the degree of
Doctor of Philosophy in Statistics

College of Science and Engineering
School of Mathematics and Statistics
December 2021

Abstract

In this dissertation I present developments in statistical methodologies that deal with interdependent data, i.e. data in which the units of observation are connected to each other resulting in a network of interdependence between them. Data considered interdependent poses a challenge to traditional statistical methodologies that assume units of observation to be independent and identically distributed. I focus on networks, and in particular social networks, as a tool to characterise these units of observation, called nodes, their observable attributes, and the connections between them. The developments in this dissertation are used to try to answer questions about the causal relationship between the observed variables, conditional on the network structure.

In chapter 3 I present a causal analysis of the the Sexually Transmitted infections And Sexual Health (STASH) intervention and find that it had a positive effect of treatment (direct effect), but no effect of interference (effect of treatment spilling over to other individuals). I consider the methodology developed by Forastiere et al. (2020), as well as a flexible regression approach, to model the potential outcomes of the intervention for different levels of treatment and spillover, conditional on the joint propensity to be treated, directly and indirectly. Using a simulation study, I find that the proposed flexible approach has similar performance in terms of bias and uncertainty to the approach by Forastiere et al. (2020) when estimating the effect of the intervention, without the need for full information on the outcome model. In addition, our simulations suggest that regardless of methodology, estimation using a small sample produces larger uncertainty bounds.

In chapter 4 I present a methodology to identify social influence and separate it from the effect of prior similarity in bipartite event cascades, when analysed using the relational event model (REM). The REM can be used to analyse the interdependent nature of data where the behaviour by an actor can be caused by the recent behaviour of similar actors (social influence). Homophily statistics can test for such contagion, given one or more actor attributes or network relations. However, social influence along the cascade, and independent but similar behaviour as a consequence of shared attributes, are generally confounded. Using Monte Carlo simulations, I show the limits of a randomisation test as a tool to distinguish from these two competing mechanisms (influence and prior similarities). The simulations, as well as an empirical example in political science, delineate the scope conditions of the randomisation inference test used and demonstrate its efficacy under different mixture regimes of influence and similarity.

Chapter 5 presents a Bayesian methodology to estimate parameters for social networks using the exponential family of distributions via a network sampler that produces candidates in which both the connections between the nodes and their attributes are considered endogenous. Parameter estimation for networks with the exponential family is based on sampling networks candidates conditional on a fixed value of the parameter. Traditional estimation produces networks where only the connections between the nodes are switched to produce viable candidates. Fellows and Handcock (2012) developed a sampler that produces networks where both the connections *and* some nodal attributes are switched (toggled, as it is referred to in the literature) in order to generate viable samples. I propose using a Bayesian estimation routine with a sampler that also toggles node attributes and network connections, based on Caimo and Friel (2011)'s approach, to replace estimation using maximum likelihood, and produce samples from the posterior distribution for the parameter. This results in an estimating methodology that considers a data generating process in which networks are generated by changing edges and node attributes, and conditional on having a proper model, is less prone to produce degenerate results.

Declaration

I have prepared this thesis myself; no section of it has been submitted previously as part of any application for a degree. I carried out the work reported in it, except where otherwise stated. The work presented in Chapter 3 is jointly authored with Dr. Erica E. M. Moodie and Dr. Nema Dean. I delivered a condensed version of this paper as a presentation at the Berlin Epidemiological Methods Coloquium in 2021. The work presented in Chapter 4 is jointly authored with Dr. Philip Leifeld and Dr. Laurence Brandenberger, and is currently in submission. I presented a earlier version of the contents of this chapter at the Sunbelt Conference in Montreal in 2019 and the American Political Science Association Conference in 2021.

Thesis Supervisor: Dr. Nema Dean
Title: Senior Lecturer - University of Glasgow

Thesis Supervisor: Dr. Philip Leifeld
Title: Professor - University of Essex

Thesis Supervisor: Dr. Mark McCann
Title: Research Fellow - University of Glasgow - Social and Public Health Sciences Unit

Thesis Supervisor: Dr. Erica E. M. Moodie
Title: Professor - McGill University

Acknowledgments

Estoy infinitamente agradecido por todo el apoyo que he recibido de mis padres, Luz Marina y Jaime, que desde el comienzo me han incentivado a perseguir mis sueños. No estaría donde estoy sin su apoyo incondicional. Agradezco también el apoyo incondicional de mi abuela, Mary, que siempre me enseño a trabajar para conseguir lo que uno sueña. Juan, gracias por mostrarme siempre el camino a la excelencia. To the rest of my family, Manu, Barbara, Lily: thank you for keeping me fed, warm and jolly throughout these years. Kevin, Tim, Jovana, Jill, Daniel, Ben, Eva and Titi: the support I have received from you has kept me going, even in the coldest of winters and the darkest of times. Fritz, your wise advice has saved me in this journey more than once. I would like to thank the Lord Kelvin/Adam Smith interdisciplinary program at the University of Glasgow for the financial support to make this project possible, as well as to the Early Career Mobility Scheme for allowing me to do a research stay at McGill University. I could not imagine a better team of supervisors than Dr. Nema Dean, Dr. Philip Leifeld, Dr. Mark McCann, and Dr. Erica Moodie. I am the most fortunate individual to be able to learn from them, academically and personally, throuhgout all these years. This project would be not exist without them. Thank you to the infinite number of individuals on the internet who have helped me fix a problem with my computer. Lupe, thank you for taking me out for a walk when I needed it the most. Rieks, words do not cut it, but without you, this would all have been impossible. With you, it has been life itself. Thank you for being my rock, my base, my home, my future, and my journey.

To Élio and Noa.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In practice, individuals, and units of observation more generally, rarely exist independently from one another. These observations are often assumed to be independent draws from a particular probability distribution. What happens when they are not? The overarching topic covering my research is the relationship between between units of observation, their observed attributes, and how they connect to one another. The primary research question that I would like to answer in this dissertation is "how to determine the causal relationship between different variables with interdependent observations?". “Causal relationship” makes reference to the fact that we want to understand how changing one variable changes another one. “Interdependent observations” means that the units of observations I will consider in my analyses are connected to each other, that their attributes are somehow interlinked, and that the causal relationship between two variables is connected to the relationship between units of observations and their characteristics.

My dissertation is mainly motivated by two questions. In the context of empirical studies where we have information about individuals and the relationships between them, can we identify whether a change in the observed attributes of an individual happened because of changes in the attributes of someone else in the network? Can we identify a change in the connections between the individuals because of a change in their attributes? The first of these questions suggests that changes in attributes can be transmitted from one individual to others through the interactions that they have, a process known as social influence. The second implies that the connections formed by individuals are determined by the observable attributes they have, a process known as social selection (Leenders, 1997).

To approach these questions I will look at three specific methodologies in the field of statistics known as social network analysis. A network is a way to represent units of observation and the connections between them. When these units are individuals, we refer to these representations as social networks. The fact that people are connected to each other means that their observable characteristics are also somehow connected. I am interested in the causal relationships between these characteristics (and not just their association), in the context of this interconnectedness.

Networks allow social scientists to explore phenomena in which connections between individuals are locally emergent, are influenced by several factors (several of them occurring at the same time), but ultimately, are structured and stochastic (Lusher et al., 2012). Lusher et al. referred to the analysis of social networks using the exponential random graph model (properly introduced in Section 2), because of the broad and rich mathematical background of the exponential family of distributions allows it to be a great candidate to investigate all sorts of situations in network science.

As such, I present developments in statistical methodologies that deal with data that are interdependent, and which we would like to use to answer causal questions. These are divided into three main chapters. The first (Chapter 3) considers interdependence in the field of causal inference in the form of the treatment to one individual affecting the outcome of a different one. The second (Chapter 4) looks at a randomisation inference-based methodology to attempt at disentangling the effect of social influence from similar behaviours in sequences of events. The third (Chapter 5) introduces a Bayesian approach to parameter estimation of social networks using the exponential random network model. In this introduction I expose the general motivation for all three projects, and how they are connected to each other.

Causal inference relates to the use of theory and methodology to estimate the effect of an event or a decision on an outcome of interest. As an independent field of research, it has relied on developments from statistics, econometrics and probability, as it was initially developed for applications in policy interventions and other types of program evaluations. Progress in the field has come with formalisations on language, notation, and more broadly, what it means to think in terms of causal inference. Some notable examples include Rubin (1974), Pearl (2009), Angrist and Pischke (2009), Rosenbaum (2010), and most recently Hernán and Robins (2020) and Cunningham (2021).

The main question causal inference looks to answer is "what would have happened to the outcome of one unit if it received treatment A instead of treatment B?" To be more precise, and looking to frame causal inference within the topic of social network analysis, this question should read "what would have happened to the outcome of one unit if it received treatment A instead of treatment B, considering that a second unit connected to the first one received a specified treatment?"

There are many contexts in which we can apply this question. One of the of the most widely used examples in the literature considers the action of giving an aspirin to someone who has a headache. Causal inference is concerned with estimating the effect of this action, even though we can never know what would have happened to that individual if we had not given them the aspirin.

There are many different possibilities for estimating causal effects, but one of the best is to find as many people that are similar to the to-be intervened individual, and randomly give the medication to some, and a placebo to others. This is known in the literature as a randomised control trial (RCT), and is considered the gold standard in causal inference research (Rubin, 1974). There are other methodological alternatives for when we have no control over who gets assigned to treatment or not. The literature commonly refers to these as quasi-experimental studies, or observational studies, the former being a subset of the latter.

When intervening on a collection of units (like individuals or regions), it is common to assume that the treatment applied to one of them does not affect the outcome of the others; this assumption is a common one in causal inference. Formally, this is referred to as there being no *interference* between units because they are considered independent from each other.

These are some examples which this assumption holds: in measuring the effect of an intravenous treatment on a group of patients, or of a legislation that is applied to one specific geographic area. In both of these examples it is not plausible for the treatment to spill over to other units: intravenous treatments cannot be shared, and a legislation applied to one country does not apply in another one. However, it is easy to think of situations where the assumption of no-interference does not hold: in measuring the test performance of a group of students where some receive the answers beforehand, and might share them with their friends, the effect of polluting a river which affecting the quality of the water down the stream, or the impact a virus that travels through the air has on a population.

Causal inference in the presence of interference is a particular subfield of causal inference research where the assumption of 'no interference' is relaxed. Situations where the units of observation are inherently linked to each other, situations suited for network analysis and in particular social network analysis, meet this criteria and hence provide an interesting range of applications with high potential for impact.

Social network analysis is a natural fit for the study of causal inference in the presence of interference. The methods developed to understand networks provide a rich background to explore how individuals interact with each other, and hence, how the treatment to some might affect the outcome of others. More specifically, I am interested in measuring the causal direct effect of a treatment, as well as the causal effect of the spillover of the treatment, i.e. when the treatment to one unit affects the outcome of another one. Developments like network-informed matching and a special version of randomised control trials allow us to explore these questions from a experimental design point of view (Hudgens and Halloran, 2008; Aronow and Samii, 2017). Dealing with observational data requires some more refined methods. In Chapter 3, I focus on the methodology developed by Forastiere et al. (2020) for observational data and show that a flexible regression approach produces similar results without the need for the kind of perfect information required by the authors.

The methodology is put to the test using data from an intervention developed by the Social and Public Health Sciences Unit from the University of Glasgow. In the Sexually Transmitted infections And Sexual Health (STASH) project, selected peer supporters were tasked with sharing educational content with their peers. Using our flexible approach, we show (and corroborate the finding) that the intervention did not work as intended and the imparted knowledge did not spread.

In Chapter 4, I change focus to a randomisation based approach to measure whether temporal order matters in the way actors connect to a set of behaviours. I take the ideas on using randomisation inference presented by Malang et al. (2019), and extend them to show that the effective implementation of this methodology is limited by the some specific characteristics of available data. Chapter 4 includes an empirical example where countries ratify environmental treaties at different points in time and find that there is evidence that the order in which they decide to ratify the treaties matters, conditional on their political and geographic characteristics.

In Chapter 5, I present a methodology that estimates parameters for a particular kind of model used to analyse social networks with the exponential family of distributions. In this model, we observe a given network and consider its characteristics, which include node attributes and the connections between the nodes in the form of an adjacency matrix. The aim of this model is to characterise the network through a set of statistics calculated on the these attributes and connections.

A likely output from this analysis are properties of the data generating process (DGP) that produced the observed network. To determine how likely observing one network is, we need to see what is the likelihood of sampling that particular one from all the different possibilities given the data generating process. A common procedure to sample networks that come from the same DGP is to use a Markov chain Monte Carlo process to generate several draws.

The sampled networks have different network structures, i.e. a unique configuration of connections that reflect network attributes. Up to this point, most computational sampling implementations rely on software that only changes the connections in the network to produce new network candidates, while maintaining the node attributes as a given. This speaks to a larger issue in network inference: the data generating process only considers networks with different configurations, but the same node attributes. Alternatively, models like the network autocorrelation model try to estimate data generating processes for node attributes, given network structure.

Fellows and Handcock (2012) proposed a sampling algorithm that allows for network with different network structures and different node attribute values, all stemming from the same data generating process. The innovation presented in this dissertation couples that same sampler with a parameter proposal algorithm that uses Bayesian inference and reduces the possibilities generating unviable network samples.

The rest of this dissertation is organised as follows. Chapter 2 introduces the statistical background required for all of the presented developments. Chapter 3 presents the developments in the field of causal inference in the presence of interference and the empirical application. Chapter 4 shows the developments related to the limits of a randomisation inference methodology, as well as the practical example, while Chapter 5 presents the two algorithms used to sample networks with toggling of both network connections and node attributes as well as the Bayesian approach to estimation. Chapter 6 summarises the results from the three core methodological chapters and presents some closing remarks.

Chapter 2

Statistical Background

2.1 Statistical modelling

I begin this subsection with a small introduction to some notation and ideas behind inferential statistics. The field of statistical inference is devoted to developing statistical tools to infer or provide insight about a variable of interest using only observable information. This procedure of 'belief' (Young and Smith, 2010; Rougier, 2017) is present in different areas of science. However, what makes the approach in statistics unique is the use of probability theory to quantify such belief.

A statistical model can be described as a device statisticians use to connect the things we would like to know, which are often unobservable, with the things we are able to measure. We assume that these two collections (the observable and the unobservable) are quantifiable, or in other words, that they are numerical. This is equivalent to saying that they can both be taken from the same distribution. They are also random variables, which means that before being measured, we do not really know their actual value.

I will now introduce some important notation. Let $\boldsymbol{X} = \{X_1, X_2, \cdots X_n\}$ be the set of observable quantities we take as input for our statistical model. $\boldsymbol{X}$ is the set of all possible outcomes for $\boldsymbol{X}$ - it can be referred to as the *realm* of $\boldsymbol{X}$. Considering that statistical models take the form of a family of probability distributions, we can define the complete set of probability distributions for $\boldsymbol{X}$ as

$$\mathcal{P} = \left\{ \mathbf{p} = (p_1, \cdots, p_k) \in \mathbb{R}^k : \ \mathbf{p} \geq 0, \sum_{i=1}^{k} p_i = 1 \right\},$$

where $p_i = \mathbb{P}(X = x_i)$, and represents the probability of observing the random variable X to be equal to x_i, and $k = ||\mathcal{X}||$ refers to the number of elements in $\mathcal{X}$. A family of distributions is a subset $\mathcal{F}$ of the *complete* family of distributions, $\mathcal{F} \subset \mathcal{P}$. The subset is used by the analyst to rule out which probability distributions are not relevant for the statistical model being built. This family is denoted a probability mass function in the case of countable $\mathcal{X}$, and a probability density function in the case of uncountable $\mathcal{X}$, and they depend on a parameter $\boldsymbol{\theta}$ and a parameter space Ω. Formally,

$$\mathcal{F} = \left\{ \mathbf{p} \in \mathcal{P} : \ \forall i, \ p_i = f_X(x^{(i)}; \theta) \text{ for some } \theta \in \Omega \right\}.$$

$\mathcal{F}$ is the basis for any statistical model. A sufficient statistic is an important part of many statistical models used by statisticians to reduce the amount, but not the quality, of information needed to perform statistical inference. Let $f_\theta(x)$ be the probability density function for x given θ as a parameter. We define T to be a **sufficient statistic** for θ if and only if two non-negative functions h and g can be found such that

$$f_\theta(x) = h(x)g(\theta; T(x)). \tag{2.1}$$

When T is a sufficient statistic for θ, we are able to separate f into h, which does not depend on θ, and g, which depends on θ but only through T. In other words, all the information we can get about θ through x, we can get through $T(x)$.

With these definitions we can move on to define a statistical model in a more formal manner. As Rougier (2017) suggests, models are a simplification used by researchers to try and understand a particular event. Because of this, models are subject to bias and require transparency so that readers using the model can understand what assumptions are being made. Some of the most basic models are parametric models, in which the probability distribution of interest is said to have a common mathematical form indexed by a set of parameters. As with the observations X, the parameters θ 'live' in the parameter space Ω (by which we mean that θ is contained in Ω). Formally, a model can be expressed using the triple

$$\mathcal{M} = \{\mathcal{X}, \Omega, f_X\}.$$

2.2 Social network analysis

In this section I present the background of statistical methodology for the analysis of networks, with a particular focus on social networks. A network is a representation of a set of units of observation and the connections between them. In the context of network science, we refer to these units as nodes. Social networks are networks where the nodes consist of individuals and the connections between the nodes are relationships that exist between those individuals. These can be friendships, partnerships, business dealings, or almost any representation of a link between two people or organisations. In this section I will expand on how to describe networks statistically, as well as a set of methodologies designed for making inferential claims around networks.

2.2.1 Networks

Here we look at the formal definitions of graphs and networks. It is assumed here that a graph is a certain representation form of a network, in the form of a list of the connections between the nodes. A network in turn is a social or a natural phenomenon in which relations are the primary unit of interest. However, these two terms can be used interchangeably in this and other chapters of the dissertation. Another possible representation of the connections between nodes is a matrix referred to as the adjacency matrix. The rows and columns of the adjacency matrix correspond to the nodes in the network, a 1 in the i-th row and j-th column means that the i-th node is connected to the j-th node, and 0 that they are not.

In many analyses (Hunter et al., 2008; Morris et al., 2008; Lusher et al., 2012), it is assumed that in the data generating process behind the creation of a network, the connections between nodes are a realisation of a random variable, while the attributes of those nodes are taken as a given. This assumption suggests that when considering a population, only the connections between them can change, but not their observed attributes. This implication can be observed in the estimation procedure of the `statnet` suite of packages for estimation of social networks (Krivitsky et al., 2020) (explained in detail later in Subsection 2.2.2), where the samples used to approximate the maximum likelihood estimator are limited by only proposing new networks with a different configuration of connections, but the same attribute values. Fellows and Handcock (2012) developed a new network sampler that allows for nodes to have different attributes that follows the same maximum likelihood estimation procedure.

Methods developed for the study of social networks include in their toolset ways to account for the connections between units of observation (Butts, 2001; Desmarais and Cranmer, 2012). Traditional linear models assume that individuals are not connected to each other or that the connections are somehow static and taken as a given. According to Ogburn (2017), models that do not include the possible interaction between the observations are biased at best and wrong at worst, since these are rarely observed isolated from one another. There is usually a web of connections that can produce an underlying structure of interdependence affecting the attributes and outcomes of the individuals under observation, and hence the results from estimation (Ogburn et al., 2020). Using the tools specifically designed for social network analysis allows researchers to question findings from traditional statistical methods that do not account for these interdependencies. Additionally, as high performance computational tools have become more widely available, the field of social network analysis has considerably expanded considerably. See the change from the more foundational Carrington et al. (2005) to the more practical introductions to the literature in Jackson (2010) and Lusher et al. (2012), as well as Krivitsky et al. (2020). Cranmer et al. (2017a) includes additional information on the different approaches to analyse networked data, and Kolaczyk (2017) provides a clear exposition of the (then) frontier of statistics and network analysis.

I now introduce several concepts that we are going to use throughout the entire document.

Notation

This subsection includes a set of notation conventions, definitions and concepts used throughout the graph and network literature.

A **network** G, is composed of a set of n nodes, and a set of edges, or connections between those nodes. The set of n nodes is defined using the letter N. All the connections between the nodes in a network are represented by a random variable Y, where y is a realisation of that random variable. An element of y is represented by $(i, j) \in y$, indicating a link between nodes i and j, and is called an edge. The edges can be either **weighted** (links have different relative importance in the network) or **unweighted** (all links carry the same importance); and be **directed** (a connection between node i and j does not imply a connection between node j and node i) or **undirected** (a connection between node i and j is the same as a connection between node j and node i).

The **adjacency matrix** A associated with G is an $n \times n$ matrix, a random matrix, and contains the information of the connections between nodes in G. a is a realization of A. In the unweighted setting, if two elements i and j in G are connected (i.e. if the link $(i, j) \in Y$), $a_{i,j} = 1$, and if they are not, $a_{i,j} = 0$. An undirected network's adjacency matrix is represented by a symmetric adjacency matrix. A directed network's adjacency matrix is not necessarily symmetric.

Nodes, N, in the network have **attributes**. We denote these attributes with the letter X and say that X_i^j is the j-th attribute from the i-th node. With this in mind, for all of this dissertation I will represent a network G with corresponding attributes as $G = (Y, X)$.

The following set of definitions will be useful for this dissertation.

- A **dyad** is a pair of nodes i and j in G, connected or unconnected.
- **X**, a matrix with n rows and k columns, representing the values of the k attributes for each of the n nodes in the network.
- A **walk** from node i_1 to node i_k is a sequence of nodes $\{i_1, i_2, \cdots, i_k\}$, defining a sequence of edges $\{(i_1, i_2), (i_2, i_3), \cdots, (i_{k-1}, i_k)\}$ such that $(j-1, j) \in y,\ \forall j \leq k$.
- A **path** is a walk $(i_1, \cdots, i_k)$ where each i_j is distinct.
- A **cycle** is a walk that ends in the same node, i.e. $i_1 = i_k$.
- A **geodesic** is the shortest path in terms of number of edges on a walk between two nodes.
- Y is said to be **connected** if there exists a path between any pair of nodes in the network.
- It is said that a directed graph G is **strongly connected** if every node is reachable from every other one following the direction of G. This is, for every pair of distinct nodes i and j there exists a directed path that connects them.
- Alternatively, it is said that a directed graph G is **weakly connected** if when considering it as an undirected graph it is connected (i.e. ignoring direction of edges from the original network).
- An **induced subgraph**, often referred to as just s subgraph, $G' = (Y', X') \subset G$ is a subset of the nodes in a network, with all of the edges from the original network linking these nodes.
- A **component** $G' = (Y', X') \subset G$ of a graph is a maximal connected subgraph such that

  - (Y', X') is connected

  - $i \in G'$ and $(i, j) \in Y \Rightarrow j \in N'$ and $(i, j) \in Y'$.

  Individual nodes are also components.

- The **neighbourhood** of i: $\mathcal{N}_i(y) = \{j \mid (i, j) \in y\}$ when G is an undirected network. When G is a directed network, we define the **incoming neighbourhood** of i as the set of all incoming connections to i; an **outgoing neighbourhood** is the set of all outgoing connections from i. Figure 2-1 shows a diagram of the neighbourhood around node A for a network $G = (Y, X)$, where diagram is defined as all nodes within one edge of the main node.

- The **degree** of i: $\text{degree}_i = |\mathcal{N}_i(y)|$, the number of nodes connected to i, with either incoming or outgoing edges. The **in-degree** of a node is the number of edges coming into it. Its **out-degree** is the number edges coming out of it.

- The **average degree** of G, d, is the average number of connections a node has in the network; i.e. $\sum_{i=1}^{n} \frac{d_i}{n}$.

- The **diameter** of a graph G is the longest geodesic. If G is disconnected, the diameter represents the longest geodesic in one of the components of G

- The notion of **centrality** in a network can be described as a measure of the importance of a node in the network. There are different measures, describing different types of connection or risk. Some examples include degree centrality (defined as $\text{degree}(i)/n$) and closeness centrality (the average of the shortest path length from the node to every other node in the network).

- A **graph model** is a probability distribution over the space of all graphs.

- A **bipartite network** (also referred to as a two-mode network) is a network where the nodes are classified into two separate categories or modes, and nodes from the first mode only connect with nodes from the second one. A **sender**, s is an element of the first mode, that creates a connection to the elements of the second mode, the **receivers**, r.

- An **event** is a tuple $e = (s, r, t)$ that comprises a sender s (an actor, country etc.), a receiver r (a behaviour, activity, treaty etc.), and time t. An **event sequence** is an ordered collection of events.

- A **bipartite event cascade** is a bipartite network where the order in which the connections between the two modes happened is known, and hence can be also represented as an event sequence.
- In social network analysis, the term **homophily** refers to the tendency of nodes in a network with similar attributes to one another to form ties with each other, preferentially to anyone else in the network. This is sometimes called selection. The term **influence** refers to the fact that an observable attribute of a node can be because of the influence other nodes exert on it.

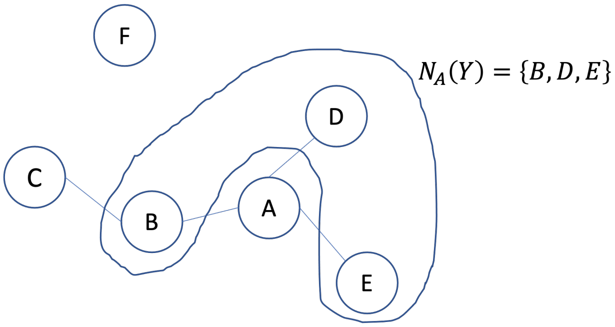


Figure 2-1: Diagram of neighbourhood of A, which comprises nodes B, D, and E. In this diagram neighbourhood is defined as all the nodes that are within one edge of the main node.

See Robins et al. (2007), Kolaczyk (2009), Jackson (2010), Butts (2011) and Lusher et al. (2012) for more information on these definitions.

2.2.2 The exponential random graph model

In this subsection I use the definitions previously introduced to define the concept of an exponential family, random networks, and the model associated with them, the exponential random network model, or ERGM. The exposition in this chapter is particularly important for Chapter 5.

The use of probability distributions to model connections between nodes stems from (Gilbert, 1959; Erdős and Rényi, 1960) The exponential family of distributions was first proposed by Holland and Leinhardt (1981), looking to model the pattern of relationships between the nodes in a network rather than the distribution of the observable attributes by those nodes. The motivation to use exponential families came from the fact that the this family of distributions tied the value of a sufficient statistic used to explain particular features exhibited by those relationships, with the estimation of a parameter for that statistic. Here I present the exponential family, and connect its usage with the concept of sufficient statistics.

First we need to define the **exponential family** distribution in the network context as a probability distribution that admits the following canonical decomposition (Shalizi and Rinaldo, 2013; Geyer, 2021):

$$p_\theta(Y = y) \propto e^{(T(y)\cdot\theta - c(\theta))}, \tag{2.2}$$

where $T(y)$ is a vector of length d with sufficient statistics that explains the object we are trying to understand, in this case Y, and, θ is the vector of d parameters (sometimes referred to as the canonical parameter) associated with those sufficient statistics. $T(y) \cdot \theta$ is the dot product between $T(y)$ and θ:

$$T(y) \cdot \theta = \sum_{i=1}^{d} T_i(y) \cdot \theta_i.$$

$c(\theta)$ is the log-normalizer, normalizer, or more formally, the **cumulant generating function**. Given $\mathcal{Y}$, the sample space as previously defined, $c(\theta)$ must be such that it represents all of the possibilities of organising a network given a set number of nodes and a vector of node attributes:

$$e^{c(\theta)} = \sum_{y \in \mathcal{Y}} e^{T(y)\cdot\theta},$$

since, $\mathcal{Y}$ is countable, as there is a limit to the number of combinations of connections between nodes. We can therefore write $p_\theta(y)$ as

$$p(Y = y|\theta) = p_\theta(Y = y) = \frac{e^{T(y)\theta}}{\sum\limits_{y' \in \mathcal{Y}} e^{T(y')\theta}}. \tag{2.3}$$

The Fisher-Koopman-Pitman-Darmois theorem (Geyer, 2021), states that for smooth nowhere-vanishing probability densities, a finite dimensional sufficient statistic exists if and only if the densities are from an exponential family. Fisher (1922), then Darmois in 1935, and both Koopman and Pitman in 1936, state: when we are dealing with exponential family distributions we can always find a sufficient statistics that would make parameter estimation possible with limited information (Daum, 1986; Brown, 1986; Nielsen and Garcia, 2009). Some further generalisations have been subsequently introduced: an example being Barankin and Maitra (1963) who proved that the probability densities do not need to be the same for the theorem to work.

I now present a couple of examples to show how the theorem applies to different distributions from the exponential family (Geyer, 2021):

- A good first example is looking at the Poisson Family. X is defined as a Poisson random variable with mean λ draws from the non-negative integers $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$ according to the following formula

  $$\Pr_{\lambda}\{X = x\} = p_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x \in \mathbb{Z}_+$$

  $p_\lambda(x)$ can be rewritten as

  $$p_\lambda(x) = \frac{1}{x!}e^{-\lambda}\lambda^x,$$

  which allows for the canonical decomposition from Equation 2.2 and the sufficient statistics definition from Equation 2.1:

  · $T(x) = x$ is the sufficient statistic.

  · $\theta = \log(\lambda)$ is the natural parameter, which has inverse function $\lambda = e^\theta$.

  · $c(\theta) = \lambda = e^\theta$ is the cumulant generating function.

  · $1/x!$ is $h(x)$ in Equation 2.1.

- Random graph: Extending this example to the realm of graphs, a good way to understand the application of sufficient statistics and exponential families is looking at the random graph model (Geyer, 2021):

  Let $G = (Y, X, \gamma)$ be a graph with $n = |N|$ nodes connected at random with probability γ, independently from other edges, and adjacency matrix A. Consider the following statistic: $T(a) = \sum_{i,j} a_{ij}$, which counts the number of edges in G. Assuming the model in which we look at the probability of an edge, $p(a|\gamma)$ would be

  $$p(a|\gamma) = \prod_{i=1}^{n}\prod_{j=1}^{n}\gamma^{a_{ij}}(1-\gamma)^{1-a_{ij}}$$

because the random connections between individuals are binomially distributed. Using the logarithm of this likelihood,

$$p(a|\gamma) = \exp\left\{\sum_{i,j}\Big(a_{ij}\log(\gamma) + (1 - a_{ij})\log(1 - \gamma)\Big)\right\}$$
$$p(a|\gamma) = \exp\left\{n^2\log(1 - \gamma) + \log\left(\frac{\gamma}{1 - \gamma}\right)\sum_{i,j} a_{ij}\right\},$$

Using this representation,

- $T(x) = x$ is the sufficient statistic.
- $\theta = \log\left(\frac{\gamma}{1-\gamma}\right)$ is the natural parameter.
- $c(\theta) = n^2\log(1 - \gamma)$ is the cumulant generating function.

which follows the formulation of the Fisher-Koopman-Pitman-Darmois theorem. This implies that $T(a) = \sum_{i,j} a_{ij}$ is a sufficient statistic for the proposed model, and that the random graph model is a full exponential family.

Having defined the concepts of graph, random variable, and statistical model, we now move ahead to define the Exponential Random Graph Model: An **Exponential Random Graph Model** (ERGM) is a statistical model set up for a network in which both its structure, and the characteristics of its nodes are used to make inferences about how and why social ties arise. Using the notation described before, a set of nodes N, associated edges y and attributes x, i.e., a graph $g = (y, x)$, is the realisation of a random variable (Y, X) with an assigned probability distribution from the exponential family.

Equation 2.3 is key in understanding the exponential random graph model (and hence important for the remainder of this dissertation), so let us spend some time explaining what it means.

We can represent a network $g = (y, x)$, so that $p(Y = y|X = x, \theta)$ represents the probability of observing the random variable Y be equal to y, conditional on the vector of parameters θ that explain the sufficient statistics assigned by the analyst to G. That probability is equal to the summary of our network through its summary statistics, $e^{T(y)\theta}$, divided by all the different possible configurations of the network available with n nodes, $\sum_{y' \in \mathcal{Y}} e^{T(y')\theta}$. So to determine how likely it is that we observed G, summarised through $T(y)$, we just need to add all the different possibilities y', through $T(y')$. Here lies the main problem with ERGM parameter estimation. There are $2^{\binom{n}{2}}$ possibilities, which becomes computationally expensive to calculate for networks with more than 20 nodes.

One possible alternative is to make the assumption that the probability of two nodes forming a connection between them is independent of any other connections in the network. This assumption makes the calculation of the probability observing the network, $p(Y = y|X = x, \theta)$, fairly straightforward, since it's the independent multiplication of all existing connections of an observed network G:

$$p(Y = y|X = x, \theta) = \prod_{i \neq j} p(y_{ij}|\mathbf{y}_{-ij}, \theta, X = x),$$

where $\mathbf{y}_{-ij}$ denotes the possible connections in the network G except for the one between i and j, and $p(y_{ij}|\mathbf{y}_{-ij}, \theta, X = x)$ refers to the probability of observing the link between nodes i and j, given the rest of the network (Strauss and Ikeda, 1990). This assumption implies that the local interactions in the networks are unaffected by global, or even regional, network behaviours. The conditional probability of observing $y_{ij} = 1$ is

$$p(y_{ij} = 1|\mathbf{y}_{-ij}, \theta, X = x) = \frac{1}{1 + \exp\left[-\theta(T(y_{ij} = 1) - T(y_{ij} = 0))\right]}.$$

$T(y_{ij} = 1)$ is the calculation of the network statistics when $y_{ij} = 1$. Following Desmarais and Cranmer (2012), to estimate θ we can use a hill climbing algorithm to find

$$\operatorname*{argmax}_{\theta} \sum_{ij} \ln\left[p(y_{ij} = 1|\mathbf{y}_{-ij}, \theta, X = x)^{y_{ij}}(1 - p(y_{ij} = 1|\mathbf{y}_{-ij}, \theta, X = x))^{(1-y_{ij})}\right].$$

This estimation procedure is called pseudo-likelihood estimation. The benefit of this estimation is that it can be calculated using a generalised linear model. One of the downsides, aside from having to use this independence assumption, is that it has been proven that the estimation is biased (Desmarais and Cranmer, 2012).

The best we can do then is an approximation that uses Markov chain Monte Carlo methods (Frank and Strauss, 1986; Hunter and Handcock, 2006). We need this approximation to find the vector of parameters that maximises the likelihood function in Equation 2.3. Following Hunter and Handcock (2006) (who in turn follow Geyer and Thompson, 1992), we are looking to calculate $\exp\left(c(\theta) - c(\theta^0)\right)$ as a function of θ, where θ^0 is fixed and known. We know that $\exp(c(\theta)) = \sum_{y' \in \mathcal{Y}} e^{T(y')\theta}$, which means that

$$\exp\{c(\theta) - c(\theta^0)\} = \frac{\displaystyle\sum_{y' \in \mathcal{Y}} \left(e^{T(y')\theta}\right)}{\displaystyle\sum_{y' \in \mathcal{Y}} \left(e^{T(y')\theta^0}\right)}$$

$$= \frac{\displaystyle\sum_{y' \in \mathcal{Y}} \left(e^{T(y')\theta}\right) \cdot \frac{e^{T(y')\theta^0}}{e^{T(y')\theta^0}}}{\sum_{y' \in \mathcal{Y}} \left(e^{T(y')\theta^0}\right)}$$

which can be rearranged as

$$= \sum_{y' \in \mathcal{Y}} \left(e^{T(y')(\theta - \theta^0)}\right) \cdot \frac{e^{T(y')\theta^0}}{\sum_{y' \in \mathcal{Y}} \left(e^{T(y')(\theta^0)}\right)}$$

and this is just the expected value over θ^0 of $e^{T(y')(\theta-\theta^0)}$

$$= \sum_{y' \in \mathcal{Y}} \left(e^{T(y')(\theta - \theta^0)}\right) \cdot \frac{e^{T(y')\theta^0}}{\sum_{y' \in \mathcal{Y}} \left(e^{T(y')(\theta^0)}\right)}$$

We can estimate this expected value using a sample mean

$$\exp\{c(\theta) - c(\theta^0)\} \approx \frac{1}{M} \sum_{m=1}^{M} e^{(\theta-\theta^0)T(y_i(m))}, \tag{2.4}$$

where $y_i(m)$ is a sample of a random graph from a distribution defined by θ^0. A larger sample ensures convergence to the desired expected value, and can be obtained via Markov chain Monte Carlo. This is the methodology used by the `statnet` ERGM estimation procedure in *R*.

The next step, then, is to get valid samples of these networks. The way to do this is by using an McMC sampler that uses a Metropolis-Hastings algorithm (Hastings, 1970). In short, the sampler is a mechanism that generates new networks conditional on a specific set of parameters. A short description of the algorithm is given here:

For a given θ^*, a set of summary statistics $T(\cdot)$, and a given number of nodes, n, we start the algorithm with a network with edges allocated at random, and refer to it as G^*:

1. We need to change one of $n \cdot (n-1)$ dyads in the network, some of which are the edges that are already present in G^*, from its current value to the opposite value. If we select to change y_{kl}, and its value is 1, we assign it to 0, otherwise we turn 0 to 1. This procedure is referred to as **toggling** the edge. This problem can be split in two, where with probability ρ we decide to toggle one of the available non-connected dyads, and with probability $1-\rho$ we decide to toggle one of the existing edges.

2. Without loss of generality, we decide to toggle one of the existing edges. Randomly selecting one to "turn on/off", produces a new network G^{**}, different from G^* only in that one edge we decided to goggle.

3. We now compare G^* and G^{**} using the vector of summary statistics $T(\cdot)$ to create the Hasting's ratio, r:

$$\pi = \frac{p_{\theta^*}(Y = G^{**})}{p_{\theta^*}(Y = G^*)} = \frac{e^{T(G^{**})\theta^*}}{e^{T(G^*)\theta^*}} \cdot \cancelto{1}{\frac{c(\theta^*)}{c(\theta^*)}}.$$

   $\frac{c(\theta^*)}{c(\theta^*)}$ cancels to one since both values represent all the possible configurations of connections, $\mathcal{Y}$, so they approximate to the same value.

4. We decide to accept the proposal network, G^{**}, when r is larger than a random uniform number between 0 and 1, and make $G^* = G^{**}$.

After a large number of iterations, this algorithm will produce a network conditional on θ^*, that we can use for the approximation described in Equation 2.4.

Notice that in step 3 of the algorithm we are required to calculate both $e^{T(G^*)\theta^*}$ and $e^{T(G^{**})\theta^*}$, which will undoubtedly extend the time required to generate sample networks. This problem is particularly noticeable when n is large, since the amount of time required for convergence increases with the size of the network. The alternative is to use the concept of "change statistics" (Morris et al., 2008).

The relationship between the two mentioned probabilities, i.e., the change in probability that selected node to toggle, $(i, j) \in E$, is $e^{(T(G^{*})-T(G^{**}))\theta^{*}}$. The key element in this equation is $T(G^{*}) - T(G^{**})$, which determines the difference in the summary statistics between the two networks, and is defined as $\Delta(G^{*}, G^{**})$. In the case of the ERGM, this difference is the change in the value of the statistics after only one of connections between two nodes is toggled. The problem of having to calculate both $e^{T(G^{*})\theta^{*}}$ and $e^{T(G^{**})\theta^{*}}$ separately and for every step of the algorithm is then reduced to calculate $e^{\Delta(T(G^{*}),T(G^{**}))\theta^{*}}$. This ratio between the proposed network and the current network is called the acceptance ratio.

Step 1 in the algorithm presented above allows some flexibility when it comes to choosing when to toggle existing edges, or just any of the dyads in the network. The popular software for estimating ERGMs, statnet, in their “tie-no tie” sampler (the default for ERGM estimation) uses a 50/50 chance of selecting from the group of edges already present in the network or a random empty dyad (see `https://rdrr.io/cran/ergm/man/ergm-proposals.html`). In the sampler developed for this dissertation, described in detail in Chapter 5, we propose using a user-defined probability of selecting an edge from the network.

2.2.3 Maximum likelihood estimation

A likelihood function models the joint density of data we are are interested in analysing, as a function of the parameters we are trying to estimate (θ, in our case). The maximum likelihood estimator, $\hat{\theta}$ is the value of θ that maximises the likelihood function (Wasserman, 2010). In order to model complex processes, exact calculation of the maximum likelihood estimate (MLE) is usually impossible, but approximation methods that use Markov chain Monte Carlo (like the one described in the previous section) are available (Geyer, 1991). Maximising the likelihood function is equivalent to maximising the log of the likelihood function, so often it is easier to proceed in this manner. The parameters that come from the maximisation represent the log-odds of adding one additional edge to the network.

Let us refer to the log-likelihood of Equation 2.3, by $\ell(\theta)$. From above we know that this is the expectation of a function of a random network, where the random behaviour is governed by θ_0 (Hunter and Handcock, 2006). The quantity $\ell(\theta) - \ell(\theta_0)$ is defined as the likelihood ratio between

$c(\theta)$ and $c(\theta_0)$. We want the θ that maximises this difference. We can approximate this quantity by using the approximation defined above, $\frac{1}{m}\sum_{i=1}^{m} e^{(\theta-\theta^0)T(y_i)}$. The difference in log-likelihoods, will converge as $m \to \infty$, assuming the Markov chain version of the strong law of large numbers (Meyn et al. (2009) (edition from 1993) in Hunter and Handcock (2006)). We will introduce a summary of how this maximisation works, but for more details please follow Hunter and Handcock (2006) for a detailed exposition on the procedure.

$\omega_{\theta_0}(\theta) \doteq \ell(\theta) - \ell(\theta_0)$ is a function of θ, with a known θ_0. We can maximise this function with respect to θ. This is, finding the value of θ that maximises the log-likelihood of Equation 2.3. Using the approximation from above (Equation 2.4)

$$\ell(\theta) - \ell(\theta_0) \approx (\theta - \theta_0)T(Y = y) - \log\left[\frac{1}{m}\sum_{i=1}^{m} e^{(\theta-\theta^0)T(y_i)}\right].$$

Choosing the right value of θ_0 is crucial for the convergence of this maximisation. A common approach used by the statnet suite of software for estimation of network parameters, is starting with the estimate from the pseudolikelihood routine.

To find the maximum of $\omega_{\theta_0}(\theta)$, we differentiate it with respect to θ, and make it equal to 0. The following calculation was made for a univariate function, but it can be extended to a vector $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$. The implication of the canonical representation used in the sufficient statistic is that the likelihood ratio only depends on θ through $T(y)$. So, if we wish to maximize the likelihood, we only need the information provided by $T(y)$. Individually looking at the derivative of the log-likehihood, $\log \ell(\theta)$,

$$\begin{aligned} \frac{d \log \ell(\theta)}{d\theta} &= 0 \\ \frac{d \log \ell(\theta)}{d\theta} &= \frac{d}{d\theta} T(y)\theta - \frac{d}{d\theta} \log c(\theta) = 0. \end{aligned}$$

We know that $T(y)$ is a sufficient statistic, which, following our definition of sufficient statistics in Section 2.1 means that it only depends on y and not on θ. So now we have that

$$\frac{d \log \ell(\theta)}{d\theta} = T(y) - \frac{d \log c(\theta)}{d\theta} = 0$$

Since $c(\theta) = \sum_{y' \in \mathcal{Y}} \left(e^{T(y')\theta} \right)$,

$$\frac{d \log c(\theta)}{d\theta} = \frac{1}{c(\theta)} \frac{d}{d\theta}(c(\theta)) = \frac{1}{c(\theta)} \sum_{y' \in \mathcal{Y}} \left(\frac{d}{d\theta} e^{T(y')\theta} \right) = \frac{1}{c(\theta)} \sum_{y' \in \mathcal{Y}} \left(T(y') e^{T(y')\theta} \right),$$

which is just

$$\sum_{y' \in \mathcal{Y}} \left(T(y') \frac{e^{T(y')\theta}}{c(\theta)} \right) = \sum_{y' \in \mathcal{Y}} \left(T(y') p_\theta(y') \right) = \mathbb{E}_\theta[T(y)].$$

When we evaluate the likelihood at the point which maximises its value, $\mathbb{E}_\theta[T(y)] = T(y)$. In the case of several θ parameters, this is equivalent to saying that the maximum likelihood estimator $\hat{\theta}$ satisfies

$$\nabla(\ell(\theta)) = \nabla(\hat{\theta})[T(y^{obs}) - E_{\hat{\theta}} T(y)] = 0,$$

as mentioned in Hunter and Handcock (2006). Hunter and Handcock propose the following methodology to update the values of θ, using the method of Fisher scoring (see Efron, 1978), which requires calculating the following value:

$$I(\theta) = \nabla(\hat{\theta})^\intercal var_{\hat{\theta}}[T(y)] \nabla(\hat{\theta}).$$

With this value, the updating of θ from θ^k to θ^{k+1} to get to the maximum likelihood estimator is done in the following way:

1. Select an initial value of θ, θ^0.

2. Generate m samples of the network conditional on θ_0.

3. Iterate using the following formula:

$$\theta^{k+1} = \theta^k + \left[I(\theta^k) \right]^{-1} \frac{d\ell(\theta)}{d\theta},$$

to obtain a maximising value of $\omega_{\theta_0}(\theta)$.

4. If the variance of $\omega_{\theta_0}(\theta)$ is relatively small, accept the proposed $\hat{\theta}$ to be the Markov chain Monte Carlo Maximum Likelihood Estimator for θ.

To get an approximation of $c(\theta)$, that is, to generate the m samples from the above description, we have to generate draws from the network random variable to estimate the probability of our observed network occurring. This means taking the estimated parameters from our assumed model $\mathcal{E}$ and using them to estimate a probability distribution. A problem researchers have come across while performing this estimation is that of degeneracy, in which the sampled networks are almost always either empty or complete (Handcock, 2003; Caimo and Friel, 2011). This generates a problem in the estimation because empty and full networks are not useful in calculating the probability of observing a one in particular (Lusher et al., 2012). There have been some attempts to solve this problem mathematically (Li, 2015), but that research and some others (Hunter et al., 2012) have suggested that the problem arises when poorly specified models are used, i.e. the set of summary statistics used does not properly explain the observed network or contain redundancies, and hence the requested samples are not valid.

Formally, following Rinaldo et al. (2009), degeneracy is a condition in the estimation of the process where

- A combination of the estimated parameters (θ) suggests only a small number of graphs have substantial non-zero probabilities of being observed;

- The estimate of θ available through Maximum Likelihood Estimation does not exist, or the estimation procedure does not converge;

- The estimate of θ would make the observed network very unlikely.

One approach to deal with degeneracy is presented by Hanneke et al. (2010), in which degeneracy is regarded as a non-linear system, depending chaotically on different parameters. Alternatively, Schweinberger and Luna (2017) developed an different approach to deal with degeneracy by exploiting the local dependency, natural to hierarchical exponential random graph models. A meticulous approximation proves to be a way of understanding it further.

2.3 Bayesian estimation of exponential random graph models

In this section I present an alternative to maximum likelihood estimation in the form of Bayesian estimation of parameters for exponential random graph models. From Section 2.2.1 we learned that the problem of getting an estimate for θ in Equation 2.3 is that we need a sample of networks we can use to estimate $c(\theta)$. The problem with this estimation routine is that, in many cases, and in particular when the set of summary statistics we propose is not adequate, the sampling procedure generates empty or full networks, or θ candidates that are not compatible with the observed network, reducing the speed of convergence of the algorithm. The literature refers to this as degeneracy. Caimo and Friel (2011) (following Koskinen (2008)) developed a methodology that produces estimates of θ using Bayesian estimation and bypasses the problem of the intractability of calculating $c(\theta)$ (see also Koskinen (2004); Koskinen et al. (2010)). Similarly to Subsection 2.2.2, the topics covered in this section are used in Chapter 5.

For this we need an introduction to Bayesian estimation.

2.3.1 Bayesian statistics

As mentioned in Section 2.1, the models we consider for statistical analysis are comprised of data and an assumption of how these variables are distributed. This is referred to as the probability model $f(Y|\theta)$, where θ is a vector of parameters, usually unknown, that determines how f is shaped. Following this, we would like to use our observations Y to say something about, θ.

Using the law of conditional probability, we can say that the probability of two events taking place $P(A \text{ and } B)$ is equal to the probability of A happening, conditional on B, times the probability of B taking place. Symmetrically, we can say that the probability of A and B is the probability of B conditional on A, times the probability of A. This leads to the equality

$$P(A|B)P(B) = P(B|A)P(A).$$

We can replace A and B with the data that we observe Y, and the parameter that describes the probability function for said data θ. The probability of observing our data Y, conditional on the observed parameter θ times the probability of observing that parameter $P(\theta)$, is equal to the probability of observing that parameter θ conditional on the data Y, times the probability of observing Y.

Formally,

$$f(\theta|Y) = \frac{f(Y|\theta)f(\theta)}{f(Y)}. \tag{2.5}$$

Equation 2.5 says that the posterior probability of θ being the parameter for the distribution function for our data Y, after we have observed Y is equal to the probability of observing Y conditional on θ, times our prior understanding of θ, divided by the marginal probability of Y, $f(Y)$. In many situations, this marginal probability is very hard to calculate, so we represent equation 2.5 as

$$f(\theta|Y) \propto f(Y|\theta)f(\theta).$$

where $f(\theta)$ represents the prior distribution of the parameter of interest. We can think about this function as everything we think we know about θ without having seen any data related to it. We will update this understanding using the observed data, to generate what is called the posterior distribution $f(\theta|Y)$. The estimate of this update can be done using the Metropolis-Hastings algorithm when a closed form solution does not exist. The M-H algorithm navigates the parameter space, and proposes new values of θ. These proposals are accepted according to a specific acceptance rule (defined later), typical of Metropolis-Hastings procedures. When the proposed parameters converge to a relatively set of stable values, the algorithm should stop and consider these samples as being from the posterior distribution. It is important that the algorithm navigates as much of the admissible parameter space to ensure that whatever posterior distribution is generated, has properly covered the space.

In recent years, the literature has been focused on improving many aspects of the methodology (Koskinen, 2008; Caimo and Friel, 2011; Thiemichen et al., 2015; Schweinberger et al., 2020). Of particular interest for this dissertation is the mixing of the proposed parameters to ensure proper navigation of the parameter space, as well as the use of auxiliary functions to deal with intractable likelihood functions.

2.3.2 Bayesian exponential random graph model

We can rewrite Equation 2.3 in terms of the Bayesian paradigm described above. This is:

$$\pi(y|\theta) = \frac{q(y|\theta)}{c(\theta)}, \tag{2.6}$$

where $q(y|\theta) = e^{T(y)\cdot\theta}$. This means that the probability of observing $Y = y$ conditional on θ is the exponential of the summary statistics times θ, divided by the normalising constant.

When performing Bayesian inference, as described in Subsection 2.3.1, we want to estimate the posterior distribution of a parameter conditional on the observed data,

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta).$$

In the case of Equation 2.6, to estimate the posterior distribution of the parameters conditional on the observed graphs we can use a Metropolis-Hastings algorithm like the one described in Subsection 2.2.2 to decide on whether to move from one value of θ to the next proposed one (θ^*) by using the following Hasting's acceptance ratio:

$$\alpha = \min\left\{1, \frac{q(y|\theta^*)\cdot\pi(\theta^*)}{q(y|\theta)\cdot\pi(\theta)} \cdot \frac{c(\theta)}{c(\theta^*)}\right\}, \tag{2.7}$$

which is the probability of accepting θ^*. The problem in network analysis, as with the maximum likelihood estimation, is the calculation of $\frac{c(\theta)}{c(\theta^*)}$, since this requires calculating two intractable values. Following Murray et al. (2012) [1], Caimo and Friel (2011) propose using an algorithm that samples from an augmented distribution. This means, instead of following

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$$

to arrive at the posterior distribution, they propose

$$\pi(\theta', y', \theta|y) \propto \pi(y|\theta)\pi(\theta)\eta(\theta'|\theta)\pi(y'|\theta'),$$

[1] originally published in the Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence in 2006

where $\eta(\theta'|\theta)$ is an arbitrary distribution for the augmented θ', that depends on θ. A possibility for $\eta()$ is a random walk distribution centred at θ. The distribution marginalised over $\theta' y'$ is the posterior distribution we are looking for. Let us explore how this algorithm proposes new values of θ.

Given a current value of θ:

1. Draw a value of θ' from $\eta(\theta'|\theta)$.
2. Using the sampled θ', draw an auxiliary sample of the network y, from the parallel distribution $\pi(y'|\theta')$.
3. We are now going to evaluate moving from the current pair (θ, y) and (θ', y'), to the exchanged pair (θ', y) and (θ, y'). This is done using the following ratio:

$$\alpha = \min\left\{1, \frac{q(y'|\theta)}{q(y|\theta)} \frac{\pi(\theta')}{\pi(\theta)} \frac{\eta(\theta|\theta')}{\eta(\theta'|\theta)} \frac{q(y|\theta')}{q(y'|\theta')} \cdot \cancelto{1}{\frac{c(\theta)c(\theta')}{c(\theta')c(\theta)}}\right\}. \tag{2.8}$$

Notice how the problem of the intractable normalising functions is removed.

This calculation makes the proposal to "offer" the auxiliary θ' to the data y, and in the same manner, to "offer" the the current parameter θ to the auxiliary data y' (Caimo and Friel, 2011). We evaluate how likely it is that y and θ' are affine to each other by using the ratio $q(y|\theta')/q(y'|\theta')$. At the same time, we measure how likely it is that y' and θ are affine to each other using the ratio $q(y'|\theta)/q(y|\theta)$.

The relationship between θ and θ' is dictated through $\eta(\cdot)$. If this is a symmetric function, a requirement that does not change the dynamics of the algorithm, Equation 2.8 can be rewritten as

$$\alpha = \min\left\{1, \frac{q(y'|\theta)}{q(y|\theta)} \frac{\pi(\theta')}{\pi(\theta)} \frac{q(y|\theta')}{q(y'|\theta')}\right\}. \tag{2.9}$$

If we compare Equation 2.9 with Equation 2.7, we see that the ratio $\frac{q(y'|\theta')}{q(y|\theta)}$ can be thought of as the importance sampling estimate of the ratio between $c(\theta)$ and $c(\theta')$. This means that the exchange algorithm replaces the part in the maximum likelihood estimation routine that proposes a new value of θ. If the proposed value of y is stable conditional on θ (similarly for the auxiliary draws), we ensure that we are approximating the true posterior distribution, guaranteed by the fact that the summary statistics are sufficient statistics of the probability model (Caimo and Friel, 2011).

2.4 Relational event model with temporal decay

In Chapter 4, I present a methodology that helps disentangle the role that order plays in a sequence of events conditional on the attributes of the agents that execute those events. To model the event sequences, I first need to introduce the concept of a relational event model, which is what I do in this section. Relational event models (REM) are survival models with network dependence across events. Survival models typically bring several elements together: The hazard rate,

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P_i(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \tag{2.10}$$

specifies the probability of an event to occur per time period, for example the probability that some unit i produces an event at any point in time T, expressed as an instantaneous rate over time increments, given that the event has not happened yet.

The survivor function,

$$S_i(t) = P_i(T > t) = 1 - F_i(t), \tag{2.11}$$

specifies the probability that an event has not occurred at time t, i. e., that unit i's time to event, T, is larger than the current time t. The survivor function can be parametrised with different functional forms (with cumulative distribution function F), and in the case of no assumed dependence on time (i. e,. the number of realized events at any time before t does not influence $S(t)$), this functional form could be exponential, such that

$$S(t) = e^{-\lambda t}, \tag{2.12}$$

which results in a constant (baseline) hazard rate $h(t)$.

The probability that an event occurs at time t is then the product of survival and hazard (in other words, the probability that an event has not occurred yet and is about to occur), and the likelihood function is the multiplication of this product for all units:

$$\mathcal{L} = \prod_{i=1}^{n} S(t_i)h(t_i), \tag{2.13}$$

assuming no censoring, defined as the partial observation or measurement of an event.t_i makes reference to the idea that if $i < j$, then t_i ocurred before t_j.

The proportional hazard model (Blossfeld and Rohwer, 2001; Butts, 2008) links covariates to the model by exponentiating a predictor term and multiplying it with the baseline hazard, $h_0(t)$:

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right), \tag{2.14}$$

where $h_0(t) = 1$ in the exponential case and $\exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right)$ represents unit-level differences that scale the baseline hazard up or down. Here $\mathbf{x}_i$ represents the vector of covariates and $\boldsymbol{\beta}$ is the parameter that explains the correlation between $\mathbf{x}_i$ and the hazard rate.

Relational event models differ from plain survival models by including sufficient statistics that capture dependencies in the sequence of past events and their effect on the probability of an event at the current time. The definition of the hazard of an event (known as the rate function) therefore changes to:

$$h(e) = \lambda(s_e, r_e, \mathbf{X}_e, E_{t-1}, \boldsymbol{\theta}) = \exp\left[\lambda_0 + \theta^\top u(s_e, r_e, \mathbf{X}_e, E_{t-1})\right], \tag{2.15}$$

now including a function u, which returns a vector of sufficient statistics over the elements of a hypothetical event e (like the senders s_e and receivers r_e), along with its set of covariates $\mathbf{X_e}$ (for senders and receivers, depending on the vector of sufficient statistics being used), the past event sequence before the current time, E_{t-1}, and what Butts (2008, 166) calls a “pacing constant” (λ_0), which acts as a baseline temporal scale. (I omit different event types, as specified in the original exposition by Butts (2008), for simplicity.)

With this definition in mind, the probability of an event e taking place is the product of its hazard and the survivor function from the time of the last observed event to the current one for all events that could have occurred, including the event and all non-events. The likelihood of the event sequence is the product of these probabilities for all events in the sequence:

$$\mathcal{L}(E_t|\mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[h(e_i) \prod_{e_i'|t_{e_i}} S\left(t_{e_i} - t_{e_{i-1}}\right) \right]. \tag{2.16}$$

This formulation permits piece-wise constant hazards with conditionally exponentially distributed waiting times from one event to the next with survivor function $S(t) = e^{-\lambda(t-t')}$, similar to Equation 2.12. (For brevity, I assume the timeline ends with the time of the last event, such that we do not have to account for the time between the last event and the end of the observation period.)

A simplified (partial-likelihood) version of the proportional hazard model and, analogously, the relational event model is the stratified Cox proportional hazard model. It can be employed if the functional form of time is unknown or erratic or if there is only information about the temporal order of events but not the exact timing. In the plain (non-relational event) case, the conditional probability of an event happening at time t_j (conditional on it not having happened before) is the hazard rate at t_j over the sum of the hazard rates of all possible events in that could have taken place at time t_j (these are contained in the **risk set**, R_j):

$$\frac{h_i(t_j)}{\sum_{l \in R_j} h_l(t_j)} = \frac{h_0(t_j) \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sum_{l \in R_j} h_0(t_j) \exp(\mathbf{x}_l^\top \boldsymbol{\beta})} = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sum_{l \in R_j} \exp(\mathbf{x}_l^\top \boldsymbol{\beta})} \tag{2.17}$$

Here, the baseline hazard is cancelled out, and right-censored events (events where the event is observed up to a specific point that not always coincides with its end, denoted by dummy variable δ below) and time points without any event are omitted. This leads to the partial likelihood

$$\mathcal{L} = \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sum_{l \in R_j} \exp(\mathbf{x}_l^\top \boldsymbol{\beta})} \right]^{1-\delta_i} \tag{2.18}$$

(omitting right-censored events and time points without any event), which resembles a conditional logit model and can be estimated accordingly. The risk set in the stratified Cox model – the set of all events that can occur at time t_j – is the equivalent of the matched set in conditional logit. For simplicity, I employ the stratified Cox model with discrete time for estimation and simulation, but the results should hold with other variants of the REM.

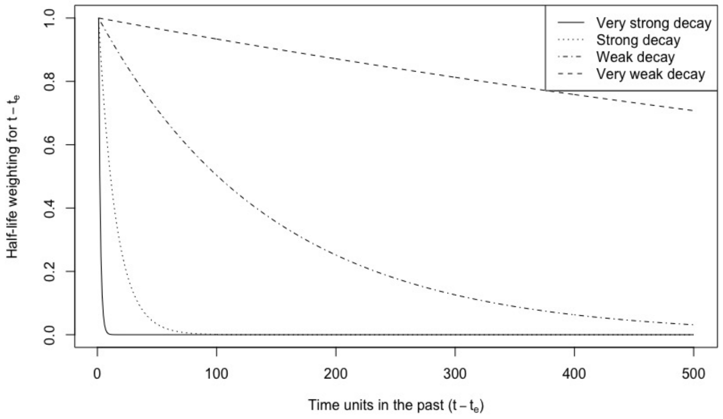


Figure 2-2: Half-life parameter. Very strong decay (solid line); Strong decay (dotted line); Weak decay (dot-dash line) ; Very weak decay (dashed line).

In the relational event model case, the analogous model can be written as

$$\mathcal{L}(E_t|\mathbf{X},\boldsymbol{\theta}) = \prod_{i=1}^{n}\left[\frac{h_{e_i}}{\sum_{l\in R_j} h_{e_l}}\right]^{1-\delta_i}, \tag{2.19}$$

and the risk set is now the set of all events and non-events that could have occurred at t_e.

In the analyses and simulations presented in the Chapter 4, we employ this discrete-time, time-ordered version of the relational event model estimated by conditional logit to keep the modelling setup as simple and general as possible. Further details are provided in Butts (2008).

I also employ temporal weighting of past events with a geometric decay in the computation of the statistics inside the u function, which was introduced by Brandes et al. (2009) and Lerner et al. (2013). Instead of using the event time t_e of past events when iterating through E_{t-1} in the u function in Equation 2.15, a weighted version of time $w(e)$ is employed. Recent events are weighted more strongly than earlier events, using a geometric decay function with half-life parameter $T_{1/2} \in (1, \infty)$, with larger values indicating less decay. The half-life parameter determines the extent to which earlier events still matter; a higher parameter means relatively higher weights are attached to early events. From Brandes et al. (2009) and Lerner et al. (2013):

$$w(e, t, T_{1/2}) = \exp\left\{-(t - t_e)\left(\frac{ln(2)}{T_{1/2}}\right)\right\}\frac{ln(2)}{T_{1/2}}. \tag{2.20}$$

Here, t refers to the current event time and t_e to some prior event time. This temporal decay is useful to limit the extent of long-range path dependence and has been employed in empirical applications (e. g., Lerner et al., 2013; Malang et al., 2019; Brandenberger, 2019). Without the temporal decay, elements of the past event sequence that occurred very long ago are as important for the hazard rate as very recent events. The temporal smoothing enables the REM to distinguish between cross-sectional network effects and temporal effects over the sequence of events, but it does not solve the issue that cross-sectional similarities can be correlated with the hazard rate. Figure 2-2 illustrates different half-life parameters for up to 500 time units in the past and shows that a large $T_{1/2}$ approximates the original REM formulation by Butts (2008) without temporal smoothing; the weight of past observations remains the same for future ones, regardless of how far away they are from each other.

I employ the following homophily statistic ξ to capture the extent to which a sender node is guided by past behaviour of other senders through a covariate for sender–sender similarities, following Malang et al. (2019):

$$\xi_e(E_t, T_{1/2}, \mathbf{a}_s) = \frac{\sum_{e^* \in E_{t-1}} [s_{e^*} \neq s_e][r_{e^*} = r_e]\,(1 - |q(s_{e^*}) - q(s_e)|)\, w(e^*, t_e, T_{1/2})}{\sum_{e^* \in E_{t-1}} [s_{e^*} \neq s_e][r_{e^*} = r_e]}, \tag{2.21}$$

where $[\cdot]$ denotes Iversen brackets, which yield 1 if the condition in the square brackets is true and 0 if false, and E_t is the set of all of the events that have ocurred up to time t. Every sender has an assigned attribute a, $q(s) \in [0, 1]$ is a function that returns a nodal covariate value for the time-invariant, continuously measured attribute a.

The homophily statistic in Equation 2.21 sums, for all past events that have the same receiver (Mode II) and a different sender (Mode I), the temporally weighted absolute similarity (i. e., one minus the absolute difference, $1 - |q(s_{e^*}) - q(s_e)|$) between the respective event's sender attribute value and the current focal event's sender attribute value, standardized over a count of past events with an identical receiver but different sender.

The statistic is calculated for the current sender in the event sequence, called the focal sender. The statistic between this focal sender and a receiver is larger if many other senders with similar covariate values as the focal sender established ties to the same receiver in the recent past. For example, consider an event sequence where senders are countries, receivers are different treaties up for ratification, and the attribute for the senders are different levels of political rights granted to its citizens (standardized between 0 and 1). The homophily statistic captures, from the perspective of the countries s_e that are considering ratifying treaties r_e, the extent to which other countries with a similar level of political rights have already ratified said treaties. Large values indicate strong homophily-based incentives, and small values indicate the absence of such cues. $\xi \in (0, \infty)$.

The inclusion of this statistic in the REM therefore operationalizes the influence exerted by "compatible" other individuals through similar recent behaviour. It can be included in empirical applications alongside other statistics, some of which were introduced by Butts (2008).

2.5 Randomisation inference

This section introduces the concept of randomisation inference, a statistical methodology designed for hypothesis testing. In short, the method considers a null hypothesis, calculates a statistic (designed to distinguish between a null and an alternative hypothesis) based on the observed data, generates a null distribution based on that hypothesis by permuting the original data, and determines whether the observed value of the statistic calculated can be considered extreme enough relative to this distribution to reject the null hypothesis. Randomisation inference was initially proposed by Fisher (1935), who stated that you can test a null hypothesis by permuting the labels of the observations. Winkler et al. (2014) framed it as "how often the difference between means would exceed the difference found without permutation" for tests of differences of means. The same idea can be extended to the hypothesis testing of different statistics.

Randomisation approaches are commonly employed to test whether estimates obtained from the observed data differ significantly from a randomized null distribution. In this dissertation, I also refer to randomisation inference as a *shuffle test*. In Chapter 4 I consider a particular application of randomisation inference, where the strategy is to break up the temporal sequence of event sequences (described in Section 2.4), through random permutation of the time stamps of observations.

The main operational steps to perform randomisation inference are:

1. Calculate a statistic with the observed data.

2. Generate many (e. g., $k = 1,000$) synthetic datasets by randomly reassigning the labels of the individual data.

3. Calculate the same statistic as in Step (1) for each synthetic dataset, and save all calculate statistic values to create an empirical null distribution.

4. Locate the original statistic in this empirical null distribution. Observe one of two outcomes: For a two-sided test:

   (a) Find the $\alpha/2$ and $1 - \alpha/2$ quantiles of the empirical distribution. If the original statistic lies between them we fail to reject the null hypothesis that the value of the statistic calculated with the original dataset is not extreme in relation to the empirical distribution.

   (b) If the original statistic is above the $1-\alpha/2$ quantile or below the $\alpha/2$ quantile, we reject the null hypothesis.

   For a one-sided (right-sided) test, we find the $1 - \alpha$ quantile of the empirical distribution. If the original statistic is above the $1 - \alpha$ quantile, we reject the null hypothesis.

Shuffle tests have been employed in different contexts for modelling networks. One prominent technique is the quadratic assignment procedure (QAP). It has been applied to multiple regression to permit unbiased hypothesis testing on relationships between correlation matrices or network relations (Krackardt, 1987; Dekker et al., 2007). In another application of shuffle tests to networks, La Fond and Neville (2010) sought to disentangle network formation from behaviour formation in co-evolutionary processes, but with panel data, assuming conditional independence, and a single kind of behaviour. Anagnostopoulos et al. (2008) employed shuffling to disentangle social influence (contagion, diffusion) from prior similarities (correlation, confounding) in logistic regression models of node behaviour regressed on the behaviour of network contacts. They specified a custom variant of a logit model that predicted whether a node was activated or not as a function of how many network contacts of the node had been activated in the past, similar to the idea of network autocorrelation models (Dittrich et al., 2017). Anagnostopoulos et al. (2008) argued that the coefficient for the number of activated network contacts (ρ in the network autocorrelation literature) was either due to social influence (being activated by the behaviour of network contacts) or confounding by potentially unobserved similarities among nodes. To distinguish influence from confounding, they posited that only influence had a temporal dimension to it and thus permuted the temporal order of the observed node activation sequence.

2.6 Causal inference

Causal inference is the study of causal relationships between variables of interest. For example, the effect an intervention had on a group of individuals. The way this intervention is designed determines the different methodologies we can use to infer the causal effect. Experiments, or experimental studies, usually refer to interventions where the researchers randomise the application of a treatment. They do this to make sure that none of the observable characteristics of the units of observation affect how treatment is applied, influencing the estimation of the effect of the intervention. Observational studies refer to interventions where the treatment was not randomised when applied to the individuals, or when there was no intervention at all, but we are still interested in the effect one variable might have over another one.

One of the most commonly known types of experiments is called a randomised control trial, or RCT. In an RCT, units of observation are randomly allocated to be either treated or not. Because of the fact that this treatment is applied with no relation to who the units are, this kind of experiment is considered to be of the highest standard for causal inference research. Observational studies, with or without an intervention, produces data that requires additional manipulation to be able to produce causal conclusions. This section introduces some basic ideas used in the study of causality and causal inference as a tool to make causal claims about the relationships between variables in data that comes from experimental and observational studies. The topics of this entire section are crucial to the developments in Chapter 3.

2.6.1 Directed acyclical graphs

In this subsection I introduce the concept of directed acyclical graphs, or DAGs, graphical representations of the causal relationship between different variables. DAGs are useful to researchers interested in causal inference because they allow a clear exposition of how two or more variables interact (or do not interact) with one another. Consider two different variables, A and B. If the value of one of them affects the value of another, we can draw an arrow from the first one to the second one, like so:

$$A \rightarrow B.$$

Causal inference requires us to think in terms of counterfactuals. That is, to consider the way in which all the possible values of A can affect B, and not just the way in which it was observed. Herein, however, lies the fundamental problem of causal inference: we are never going to be able to observe all the possible values of A and their effect on B, just what actually happened.

We define A as the parent of B, and B as the descendent of A. These diagrams are called, unsurprisingly, directed graphs (as introduced in Section 2.2), and indicate that the value of B is affected by the value of A, or that A has influence over B. If once we follow the direction of influence of the diagram, we are not able to get back to the place where we started, we call this a Directed Acyclical Graph, or DAG for short (Pearl, 2009; Shalizi, 2021). This kind of graphical representation has been used in causal inference to show the assumptions made by researchers in the expected flow of influence in a set of variables. As is the case with networks between nodes or individuals, every link, present or absent, needs to be supported by evidence or by an assumption.

Let us now consider adding one additional variable, C to the diagram. Given these three variables, a subset of the ways in which we can organise them in relation to one another is (Pearl, 2009):

- $A \rightarrow C \rightarrow B$. We refer to this as a chain that goes from A to B through C.
- $A \leftarrow C \leftarrow B$. This is a another example of a chain, with the opposite influence as above.
- $A \leftarrow C \rightarrow B$. This is defined as a fork from A and B to C.
- $A \rightarrow C \leftarrow B$. In this case, we define the relationship between A and B as colliding in C. In other words, C is a collider for A and B.

We can study the way causal information flows through these representations to aid statistical inference. An important fact to note here is that a link between A and B represents statistical information (like for example the correlation between A and B). This information goes in both directions of the graph, allowing us to explore the relationships between them. A is correlated to B is the same as B is correlated to A. However, *causal* information usually only flows in one direction (Janzing, 2007).

Regarding the four relationships from above,

- In the case of both chains, we can say that B is not independent of A, but it is independent from A, conditionally on C, $B \perp\!\!\!\perp A|C$. Another way of thinking about this is saying once we know something about C, we learn nothing new about A when observing B (or vice versa).
- Similarly in the case of the fork, following the Markov property of networks in which a variable is independent of its non-descendants given its parents (which follows from Markov assumptions in the same way that the future is independent of the past given the present (Shalizi, 2021)), $B \not\perp\!\!\!\perp A$ because they share the same parent, but conditioning on that, we have that $B \perp\!\!\!\perp A|C$.
- In the collider case, however, we have that $A \perp\!\!\!\perp B$, but that $B \not\perp\!\!\!\perp A|C$. The explanation for this is expanded below.

"Correlation does not imply causation" (see Engber, 2012) is a maxim commonly repeated when discussing statistical analyses. Statistical information in the form of a correlation can flow bidirectionally between two variables, with the causation flowing in just one direction. In some other scenarios, two variables might be related to one another because of a third, potentially unobserved variable. Causal inference uses statistical methods to formalise the flow of causality between different variables (Pearl, 2009).

Figure 2-3 shows a set of variables and explains the relationships between them in terms of statistical independence. Assume we have two factors F_1 and F_2 with causal influence on four variables, A_1, A_2, A_3, A_4 (Shalizi, 2021).

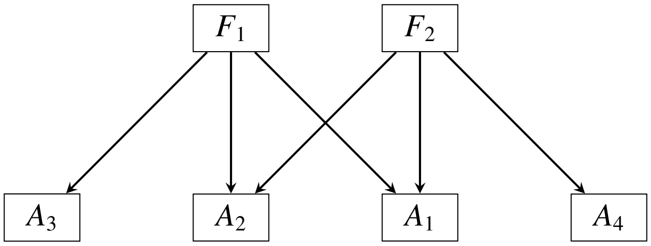


Figure 2-3: Simple Graphical Model

We would like to know what the conditional probability distribution of the variables given the factors: $p(A_1, A_2, A_3, A_4 | F_1, F_2)$. Using the facts that:

- A_1 and A_2 are unconditionally dependent through of F_1 and F_2,
- A_1 and A_3 are unconditionally dependent through of F_1. Similarly A_2 and A_4 because of F_2,
- but A_3 and A_4 are unconditionally independent,

we can say that

$$p(A_1, A_2, A_3, A_4 | F_1, F_2) = p(A_1 | F_1, F_2) \cdot p(A_2 | F_1, F_2) \cdot p(A_3 | F_1) \cdot p(A_4 | F_2).$$

Graphical models help us understand the world in terms of relations of direct dependence, and the conditional independence relations implied. In this example, the F_i variables are considered the parents of the A_i variables. In the context of graphical models we say that a variable A_i is **exogenous** if $p(A_i | A_{\text{parents}(i)}) = p(A_i)$. However, if A_3 and A_4 are unconditionally independent, conditioning by A_1 makes them dependent. This is because A_1 has information about both A_3 and A_4's parents. This is important because of how we build causal identification strategies, and what we control for.

Pearl (2009) describes "conditioning on" as "knowing the value of the variable". A variable C *blocks* a path between A and B if and only if, in a chain or fork, the middle node is in C, or in a collider, the middle node is not in C.

To understand the relation between colliders and conditional independence we can use a modified version of the example (see Figure 2-4) from Pearl (2009)'s book.

Once we know which season we are in, we can consider the fact that a tree shaker was used (to harvest pecan nuts, for example), or seeing the wind bring the leaves down to be two independent events.

However, we are interested in the sub-graph collider: Tree shaker → Floor full of leaves ← Wind. The DAG in Figure 2-4 shows that a floor full of leaves is a common effect of two conditionally independent causes, the wind or the tree shaker. If the floor is full of leaves, and we make the assumption, without loss of generality, that no one came to shake the trees, we assign a

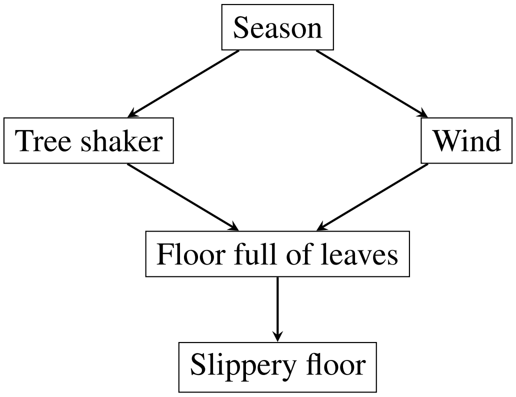


Figure 2-4: Basic DAG showing the effect of a collider on a variable

larger probability on the occurrence of strong gusts of wind. This effectively makes the two events dependent.

The situation I just described can be framed in a more general way: we are interested in the conditional distribution of a variable *B* given a set of variables grouped in *A*. If one or several variables in *A* is unconditionally independent from *B*, it should not be considered when calculating the conditional probability distribution. Including it will generate complications because the conditional distribution is going to be based on a higher number of variables than actually required, affecting converging rates and accuracy. The bias from including these variables can be reduced by using cross-validation, as presented by (Hall et al., 2004).

Once we have laid out the relationship between the variables in our model in a graphical representation, here are two ways of corroborating that we are on the right track: i) comparing the conditional independence relations implied by graphical model with the data; or ii) inferring the conditional independence assumptions from the data and determining how this translate into a possible graphical model (Shalizi, 2021).

2.6.2 Identification

Before doing any sort of statistical inference, however, we need to consider the idea of identification of our model. Given two variables of interest *A* and *B*, and additional potential variables that might affect them both, 'identification', or an identification strategy, is the process of finding the effect *A* has on *B*, clear from the influence of other variables that might affect *B*, or how *A* relates to

B. This approach will indicate whether we can arrive at the true value of a parameter with infinite observations, or if we cannot. In a way, identification determines what we can learn from infinite samples, while statistics tell us what we can learn from a finite sample (Keele, 2015).

Hernán and Robins (2020) indicate that observational studies require us to make some additional assumptions in order to be able to reach causal conclusions. Consider the case where we want to understand the effect of A on B, but A is not allocated in a way that is independent of all other variables, but is related to B via a set of covariates, L. The three main requirements for a causal effect to be identified are: i) the values of A correspond to a well defined intervention; ii) the conditional probability of receiving a particular value of A depends only on L; and iii) the probability of observing a particular value of A conditional on L is greater than zero.

When interest lies in establishing a causal link between two variables, or sets of variables, especially in the context of an observational study, we need to think of the problem in terms of the identification strategy. Statistics alone will not suffice. Another way of saying this is that causality is a matter of understanding context. With this in mind, the question we are trying to answer is whether we would be able to do causal inference for our problem given infinite amounts of information (Keele, 2015).

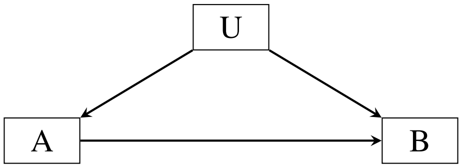


Figure 2-5: Unobserved confounding in a DAG

Consider the case presented above in Figure 2-5 in which a set of unobservable variables U is parent to both A and B, and that additionally $A \rightarrow B$. If we only observe A and B and have absolutely no information about U, $P(B|A = a)$ is going to be confounded (that is, affected through the relationship) by the presence of U. As we will see in Chapter 4, we can use Directed Acyclical Graphs to present an overview of the variables part of a statistical model, and that following a specific set of assumptions, show that we have a appropriate identification strategy. The causal identification technique used in that analysis follows the concepts from the example in Figure 2-5.

Failure to include relevant variables in the identification will result in biased or spurious statistical estimates.

According to Pearl (2009), there are three strategies we can follow to make sure that we carefully select an adequate set of control variables. These are: i) back-door criterion, or identification by conditioning; ii) front-door criterion, or identification by mechanisms; iii) instrumental variables.

The key word in this sentence is "carefully select" control variables. For the purposes of this dissertation and the strategy that we are looking to implement, we are going to look into the details of the back-door criterion.

A **back door**, as defined by Pearl (2009), is a set of variables S, relative to an ordered pair of variables (A_i, A_j), in our model, this is, $A_i \to A_j$ but not $A_j \to A_i$ if

1. no node in S is a descendant of A_i, and

2. S blocks every path between A_i and A_j that contains an arrow into A_i.

A back-door path is an *undirected* path between A_i and A_j with an arrow into A_i. These paths create confounding because of the open flow of information between the variables of interest (as we saw in previously, statistical information can flow in the opposite direction of the assumed causal flow).

Given the definitions presented above, and knowing that we are dealing with network observations, we are going to use the back-door criterion. The following is a set of rules that helps us determine whether the list of variables we are selecting to control for will meet the back-door criterion (from Entner et al. (2013)). Let $\mathcal{W}$ be the set of all the variables that do not have neither A nor B as a parent. Then,

1. If there is a set of controls S such that $A \perp\!\!\!\perp B|S$, then A has no causal effect on B.

2. If there is a $W \in \mathcal{W}$ and $S \subset \mathcal{W}$, not including W such that:

   i) $W \not\!\perp\!\!\!\perp B|S$, but

   ii) $W \perp\!\!\!\perp B|S, A$,

   then A has an effect on B, and S satisfies the back-door criterion for estimating the effect.

3. If there is a $W \in \mathcal{W}$ and $S \subset \mathcal{W}$, excluding W, such that:

    i) $W \not\!\perp\!\!\!\perp A|S$, but

    ii) $W \perp\!\!\!\perp B|S$,

    then A has no effect on B.

Situations outside of those described here will require an identification strategy that depends specifically on the causal diagram and is not going to be determined by the independence relations among the observables. Additionally, Shalizi (2021) argues that we should not control for anything which is a descendant of either B or A, that might block a directed path, activate a collider, or is just irrelevant.

I introduce the idea of conditional independence in relation to the causal relationship between the variables because of how helpful it is in terms of causal inference and interdependent data. In a footnote in Shalizi (2021), the author gives draws attention to Fowler and Christakis (2008), Christakis and Fowler (2007) and others' research suggesting certain behaviours can spread throughout a social network. Shalizi suggests that these studies rely on conditioning on the existence of a social tie between two individuals, but that this tie is actually a collider because of the latent, and hence unobservable, characteristics that determine both the outcome of a node but also its connections. As mentioned in subsection 2.6.1, activating a collider (i.e., conditioning on the variable in the collider position) creates confounding. In the other words, having latent variables in the DAG leaves a back-door open (Shalizi and Thomas, 2011).

Lerner et al. (2013) suggest that dependencies from network observations are complicated, and that past observations might not be sufficient to control for the unobservables, specially when they are not temporally close to each other. More generally, the authors suggest that conditional independence models are inappropriate as a general model to understand network evolution.

One possibility to produce correct estimators is to try to control for as many variables as possible so we can reduce the number of open back-door paths. Theorem 3.3.2 in Pearl (2009) suggests that if "a set of variables C satisfies the back-door criterion relative to (A, B), the the causal effect of A on B is identifiable". Note how this theorem does not imply that if an effect is identifiable, then the set of variables C needs to meet the back-door criterion, only that if C represents the full set of parents of A, we might be able to find the effect without controlling for all unobservables. In particular, C can include past observations of the outcome variable, in a way to control for all the information therein contained. I will employ this strategy for the developments in Chapters 3 and 4.

Another possibility to produce correct estimators in the face of confounding is relying on partial identification (Ho and Rosen, 2015) to determine bounds on the causal implications of a particular variable. Partial identification is based on the idea of using fewer assumptions when building a model to define a set of bounds to the desired estimates, instead of aiming to produce a particular point estimate. The strategy places the focus on the assumptions the researcher is using for the model: the stronger these are, the tighter the bounds. There are two possible directions to relax the model. One is on the functional form of the relations between agents, and another is on the shape of the distributions of the unobserved variables conditional on the observed variables (Ho and Rosen, 2015). An example of this approach can be found in Kang and Imbens (2016), where the authors look into the a new experimental design called "peer encouragement design". It considers network treatment effects under imperfect compliance, and define a set of estimands using partial identification based on the extent of the treatment. Similarly, Swanson et al. (2018) define partial identification for instrumental variables.

2.6.3 Potential outcomes framework

In this subsection we introduce the concept of potential outcomes. An important assumption made in traditional causal inference research is that the treatment being applied to one unit of observation only affects its own outcome, and does not affect the outcome of other units in the sample. The literature refers to this as the no-interference, or no spillover assumption, and it is a core component of as assumption made by many causal approaches: the Stable Unit Treatment Value Assumption, or SUTVA (Cox, 1958; Rubin, 1980). The Stable Unit Treatment Value Assumption has two parts. The first assumes that there are not multiple versions of treatment (also known as consistency). This assumption ensures that the potential outcomes and the realised outcomes coincide under the same treatment application. The second refers to the fact that the treatment to one unit of observation does not affect the outcome of a second one, called treatment interference. The potential outcomes framework posits that we can consider the outcome without the treatment as missing data (as opposed to data that is impossible to recover). The name "potential outcomes" suggests that an individual potentially has two outcomes, but we only directly observe one.

Consider a binary treatment, $Z = \{0, 1\}$, and an outcome variable $W(Z)$ that depends on that treatment. As mentioned above, a unit i can be either treated, $Z = 0$, or not, $Z = 1$. The potential outcomes for units i and j, that is, the possible outcomes that W can take depending on the treatment, can be represented by Table 2.1.

Table 2.1: Potential Outcomes example (reduced)

Unit	Potential outcomes		Real World	
	$Z = 0$	$Z = 1$	$Z = 0$	$Z = 1$
i	$W(0)$	$W(1)$	$W(0)$	?
j	$W(0)$	$W(1)$	?	$W(1)$

However, this can be considered an abuse of notation. Consider the following expanded version of Table 2.1, Table 2.2.

Table 2.2: Potential Outcomes example (complete)

Unit	Potential outcomes		Real World	
	$Z_i = 0, Z_j = 0$	$Z_i = 1, Z_j = 1$	$Z_i = 0, Z_j = 0$	$Z_i = 1, Z_j = 1$
i	$W_i(Z_i = 0)$	$W_i(Z_i = 1)$	$W_i(Z_i = 0)$	?
j	$W_j(Z_j = 0)$	$W_j(Z_j = 1)$	?	$W_j(Z_j = 1)$

Without the full notation present in Table 2.2, we might fail to see that Table 2.1 makes the assumption that $W_i(Z_i, Z_j) = W_i(Z_i)$. In other words, the assumption that the outcome of i only depends on the treatment to i and not the treatment to j.

One representation of the fundamental problem of causal inference is that we are never going to be able to observe the potential outcomes panels in Table 2.1. In a world where this was not the case, we could study what happens to the outcome with the observed treatment and with a different treatment strategy, i.e. the counterfactual outcome.

Let's explore the concept of potential outcomes using a guiding example. Consider a sample, $\mathcal{S}$ of size n from a larger population. $\mathcal{S} = (s_1, \ldots, s_n)$. A unit i has a set of m observable characteristics: $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,m})$, a measurable outcome variable W_i, and an assigned treatment

$Z_i = \{1, 0\}$ depending on whether it was selected into treatment or not. $\mathbf{Z} = (Z_1, \ldots, Z_n)$ represents the treatment allocations of all units of observation, and $\mathcal{Z}(n)$ is the set of all possible allocations of treatment for a sample of size n. For the simplest case of a binary treatment, this would be $\{0, 1\}^n$. Similarly $\mathbf{X}$ refers to the matrix of covariates for all units of observation.

The potential outcomes framework (Rubin, 1974, 1977, 2005) posits the existence of the outcome of interest for each unit after the application of the treatment, *and* the non-application of treatment, just as the left columns of Table 2.1 and 2.2. This approach transforms the fundamental problem of causal inference (as presented Subsection 2.6.1), into a missing data problem. The fundamental problem of causal inference is that these data will always be missing.

The treatment allocation is considered a random variable with realisation for each individual $Z_i = z_i$, and every other unit within the same sample, $\mathbf{Z}_{-i} = \{Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n\} = \mathbf{z}_{-i} = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$.

I now consider the scenario in which the treatment to one unit might have an impact on the outcome of another one. In other words, what happens when we remove the assumption of no interference between units. We can define the individual average potential outcomes (Tchetgen Tchetgen and VanderWeele, 2012; Perez-Heydrich et al., 2014; Papadogeorgou et al., 2019) for a unit i as all the possible treatment allocation combinations affecting the outcome, multiplied by the probability of observing said treatment:

$$\bar{W}_i(\mathbf{Z} = \mathbf{z}; \mathbf{X}) = \sum_{\mathbf{z}_{-i} \in \mathcal{Z}(n-1)} W_i(Z_i = z_i, \mathbf{Z}_{-i} = \mathbf{z}_{-i}) P_{\mathbf{X}}(\mathbf{Z}_{-i} = \mathbf{z}_{-i} | Z_i = z_i, \mathbf{X}), \tag{2.22}$$

where:

- $\sum\limits_{\mathbf{z}_{-i} \in \mathcal{Z}(n-1)}$: We sum $\mathbf{z}_{-i}$ over $\mathcal{Z}(n-1)$, all the different possible treatment allocations for $n-1$ observations.

- $W_i(Z_i = z_i, \mathbf{Z}_{-i} = \mathbf{z}_{-i})$: The outcome of unit i given its own treatment and given everyone else's treatment allocation according to $\mathbf{z}_{-i}$

- $P_{\mathbf{X}}(\mathbf{Z}_{-i} = \mathbf{z}_{-i} | Z_i = z_i, \mathbf{X})$: the probability of everyone else's treatment allocation determined by $\mathbf{z}_{-i}$, given that unit i's treatment is z_i, and all relevant observable covariates, $\mathbf{X}$. Notice that $P_{\mathbf{X}}(\mathbf{Z}_{-i} = \mathbf{z}_{-i} | Z_i = z_i, \mathbf{X})$ is a function of $\mathbf{X}$.

- The left hand side of the equation ($\bar{W}_i(\mathbf{Z} = \mathbf{z}; \mathbf{X})$) dependes on the entire vector of treatment allocations $\mathbf{Z} = \mathbf{z}$ and not just $Z_i = z_i$ because under treatment interference, the outcome of unit i depends not only on its specific treatment Z_i, but also on the treatment to all other units, represented by $\mathbf{Z}_{-i}$.

We assume that the treatment allocations of two individuals, Z_i and Z_j are conditionally independent given $\mathbf{X}$, when these include all the information to specify the probability of observing the treatment. This means that $P_{\mathbf{X}}(\mathbf{Z} = \mathbf{z} | \mathbf{X}) = \prod_{i=1}^{n} P_{\mathbf{X}}(Z_i = z_i | \mathbf{X})$.

Traditional statistical models assume that observations are independent and identically distributed random draws from a probability density function. This allows the probability of observing $Z_i = z_i$ given $\mathbf{X}$, the information of all the individuals the sample, $P(Z_i = z_i | \mathbf{X})$, to be equal to $P(Z_i = z_i | \mathbf{X}_i)$ the probability of observing $Z_i = z_i$ given *only* $\mathbf{X}_i$.

As we will see in later chapters, as we relax the assumption of no interference between units, we cannot assume that there is no interdependence of the observable covariates in the network, so the equality above is not straightforward. However, we assume that the interdependence between the units follows, at least in part, information contained in $\mathbf{X}$. In other words, we can include the ways in which these interactions and interdependences play out using network related variables. This suggests that, in terms of our conditional probability of treatment

$$\prod_{i=1}^{n} P_{\mathbf{X}}(Z_i = z_i | \mathbf{X}) = \prod_{i=1}^{n} P_{\mathbf{X}}(Z_i = z_i | \mathbf{X}_i), \tag{2.23}$$

where $\mathbf{X}_i$ is the vector of attributes for unit of observation i. A similar argument was made by Robins et al. (2007) explaining the Markov assumptions (initially introduced in Frank and Strauss, 1986) for network analysis in the context of the exponential random graph model. The authors argue that given the correct network terms, a pair of edges in a network that do not share a node are conditionally independent from each other.

Forastiere et al. (2020) propose a different way of estimating average potential outcomes under the presence of interference. In their approach, group average potential outcomes depend on the individual treatment and the treatment of the individuals that are in every unit's neighbourhood. The definition of neighbourhood affects the estimation, as I will present in Chapter 3. The development of that chapter depends on the theory of average treatment effects, as presented in the following subsection.

2.6.4 Average treatment effect

One possibility to determine the effect of a treatment over a group of individuals is by looking at the average treatment effect, or ATE. This is defined as the expectation of the difference of unobserved potential outcomes. The ATE is defined as:

$$\mathbb{E}[W_i(Z_i = 1) - W_i(Z_i = 0)] = \mathbb{E}[W_i(Z_i = 1) - W_i(Z_i = 0)] = \mathbb{E}[W_i(Z_i = 1)] - \mathbb{E}[W_i(Z_i = 0)],$$

The population mean for the effect of Z_i. This equation requires us knowing the potential outcomes for all units. Since this is unknowable, we are going to estimate this mean effect by splitting the ATE into parts we can actually observe. Let us assume the the proportion of the population assigned to the treatment condition is π.

$$\begin{aligned}
\text{ATE} &= \mathbb{E}[W_i(Z_i = 1) - W_i(Z_i = 0)] \\
&= \mathbb{E}[W_i(Z_i = 1) - W_i(Z_i = 0)|Z_i = 1]p(Z_i = 1) + \mathbb{E}[W_i(Z_i = 1) - W_i(Z_i = 0)|Z_i = 0]p(Z_i = 0) \\
&= \mathbb{E}[W_i(Z_i = 1) - W_i(Z_i = 0)|Z_i = 1]\pi + \mathbb{E}[W_i(Z_i = 1) - W_i(Z_i = 0)|Z_i = 0](1 - \pi) \\
&= \pi\left(\mathbb{E}[W_i(Z_i = 1)|Z_i = 1] - \mathbb{E}[W_i(Z_i = 0)|Z_i = 1]\right) \\
&+ (1 - \pi)\left(\mathbb{E}[W_i(Z_i = 1)|Z_i = 0] - \mathbb{E}[W_i(Z_i = 0)|Z_i = 0]\right) \\
&= \underbrace{\pi\left\{\mathbb{E}[W_i(Z_i = 1)|Z_i = 1] - \mathbb{E}[W_i(Z_i = 0)|Z_i = 1]\right\}}_{*} \\
&+ \underbrace{(1 - \pi)\left\{\mathbb{E}[W_i(Z_i = 1)|Z_i = 0] - \mathbb{E}[W_i(Z_i = 0)|Z_i = 0]\right\}}_{**},
\end{aligned} \tag{2.24}$$

where $W_i(Z_i = 1)$ refers to the potential outcome of unit i with treatment $Z_i = 1$, and $W_i(Z_i = 0)$ is the potential outcome of unit i with treatment $Z_i = 0$. Given that we cannot observe both of these potential outcomes, this quantity is cannot be calculated, so must estimate it. The first term in Equation 2.24 ($*$), called the average treatment on the treated (ATT), the population mean treatment effect on the units that were assigned to treatment, and the second ($**$), called the average treatment effect on the untreated (ATU), is the population mean treatment effect on the units that were not assigned to treatment (Cunningham, 2021). The following assumptions show that we can only estimate three out of the five terms introduced:

- We assume that we have a good idea of how Z is going to be distributed in the population, which means that we are able to know how many people are treated. This is, we can estimate $\mathbb{E}[Z_i]$.
- We also assume that the way treatment is assigned is independent of the outcome. Using the arithmetic properties of conditionality (the properties of the expected value conditional on a specific occurrence), we can calculate $\mathbb{E}[W_i(Z_i = 1)|T_i = 1]$ using $\mathbb{E}[W_i|Z_i = 1]$ and $\mathbb{E}[W_i(Z_i = 0)|Z_i = 0]$ using $\mathbb{E}[W_i|Z_i = 0]$.

In other words, we can estimate the expected outcome for treated observations when they were effectively treated, and the expected outcome for non-treated observations when they were not treated.

We cannot, however, estimate the expected outcome on the treated when they were in the control group, and vice versa. To proceed, we need to use the assumption that treatment is independent of the potential outcomes (which hints at why randomized control trials, or RCTs, became the gold standard for social experiments. More information on the benefits and shortcomings of RCTs can be found in Pearl, 2009). The expectation of the observed outcome conditional on $Z_i = 1$ is

$$\begin{aligned}\mathbb{E}[W_i|Z_i = 1] &= \mathbb{E}[W_i(Z_i = 0) + Z_i(W_i(Z_i = 1) - W_i(Z_i = 0))|Z_i = 1]\\ &= \mathbb{E}[W_i(Z_i = 1)|Z_i = 1]\\ &= \mathbb{E}[W_i(Z_i = 1)], \text{ which follows from the assumption of independence.}\end{aligned}$$

Assuming that treatment is independent of the outcome allows to connect the observed outcomes to the potential outcomes. With this, the average treatment effect $ATE = \mathbb{E}[W_i(Z_i = 1) - W_i(Z_i = 0)] = \mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]$. This is, “the expectation of the unobserved potential outcomes is equal to the conditional expectations of the observed outcomes conditional on treatment assignment" (Keele, 2015). The situation when the treatment is not independent from the outcome will be explored in the next section.

2.7 Propensity scores

Propensity scores were developed by Rosenbaum and Rubin as a way to reduce the complexity when comparing two groups of individuals who were assigned to a treatment in a non-experimental setting (Rosenbaum and Rubin, 1983). A propensity score is the coarsest balancing score that allows researchers to stratify a population into groups for comparison. This is, a value that ensures if two units assigned to different treatments have the same propensity score, their observable characteristics will have similar values. This is done by estimating a conditional probability model of a particular variable (usually a treatment assigned to a group of individuals) on a group of covariates that determines that treatment. Except for very special circumstances, there is no analytical formula for the propensity score, so it is usually modelled and estimated using a logit model to ensure predicted values are bounded between 0 and 1, but other functional forms are allowed (Shalizi, 2021, pp. 517). These predicted values are then used as a scalar value, called a propensity score, to compare the outcomes of units with similar values of the propensity score, but were assigned different treatments (Cunningham, 2021).

The propensity score methodology looks to estimate $P(Z_i = 1|X_i) \equiv E[Z_i = 1|X_i]$, where Z_i represents a treatment variable for unit i, and X_i is the vector of covariates for that same unit. The propensity score theorem introduced by Rosenbaum and Rubin (1983) states that if a treatment variable is conditionally independent from an potential outcomes conditional on a set of covariates, the treatment variable is also conditionally independent from the potential outcomes conditional on a scalar function of the covariates, the propensity score. Formally, $\{W_i(Z_i = 0), W_i(Z_i = 1)\} \perp\!\!\!\perp Z_i = 1|X_i$, implies that $\{W_{0,i}, W_{1,i}\} \perp\!\!\!\perp Z_i = 1|P(Z_i = 1|X_i)$. Although powerful, the theorem requires the right set of variables in X_i for the outcome to be conditionally independence of the treatment. This is, we need to assume the correct specification with respect to confounders.

There are several uses for the propensity score that help reduce confounding in causal analysis. We can use the predicted values of $P(Z_i = 1|X_i)$ to create strata that ensure the individuals inside of each one are similar in the attributes that were used to estimate the propensity score. In addition, we can use these same predicted values to do a one-to-one matching and then estimate the average treatment effect on the treated by comparing the outcome of the matched units (Angrist and Pischke, 2009). The fact that $P(Z_i|X_i)$ is a scalar reduces the dimensionality problem of trying to balance groups of treatment assignment. Furthermore, a small manipulation of the main result of the propensity score theorem, that $\{W_{0,i}, W_{1,i}\} \perp\!\!\!\perp Z_i|P(X_i)$, means we can estimate the average treatment effect $E[W_{1,i} - W_{0,i}]$ as

$$E[W_{1,i} - W_{0,i}] = E\left[\frac{W_i Z_i}{P(Z_i|X_i)} - \frac{W_i(1 - Z_i)}{1 - P(Z_i|X_i)}\right].$$

,Using the estimated propensity score as weights to adjust the value of the outcome (called inverse probability weighting) eliminates the need to match units with each other or into different groups since the weight produces a sense of relative importance in the sample. In Chapter 3 I follow Forastiere et al. (2020) in using the estimated propensity score as an explanatory variable in an outcome regression, a procedure usually referred to as covariate adjustment or regression adjustment.

2.8 Causal inference in the presence of interference

Statistical inference usually relies on the assumption that the units of observation in an analysis are independent from one another. There are several scenarios in which making this assumption is not valid, since individuals might influence each other, and this might lead to incorrect estimates of causal quantities. This is referred to in the literature as spillover or interference effects. Failing to properly consider these can lead to incorrect or misleading conclusions, as shown by Ogburn and VanderWeele (2014).

One of the first formal considerations of spillover effects in the statistics and causal inference literature is Manski's 1993 paper on endogenous social effects. In it, Manski discusses the difficulty in disentangling how an individual's outcome can be influenced by their own characteristics and by their context, where context is often described as a reference group of individuals. This difficulty in identifying the reason an individual might behave in a certain way is described by the author as a result of three scenarios: endogenous effects, exogenous (contextual) effects, and/or correlated effects. Manski indicates that correct inference is only possible if we have information about the composition of reference group, which in turn determines the influence affecting every individual.

Recently, the literature on causal inference with interference has made important progress. We rely on some of these recent developments to analyse the data from STASH (in Forsyth et al., 2018). The causal inference framework we are going to consider is that of potential outcomes, as presented in Subsection 3.2.1, initially defined by Splawa-Neyman et al. (originally from 1923; translated into English in 1990) and formalised by Rubin (1974). After the application of a simple treatment to a sample of individuals, each individual's outcome can only be observed with the treatment or without the treatment, but not both. Let us assume without loss of generality that we observe the outcome on an individual who was given the treatment.

One important assumption often invoked when considering this framework is that units of observation in a sample do not interact with one another in a way that affects their resulting outcomes (Rubin's potential outcome's framework can indeed accommodate spillover Rubin (1986)). This is, that the treatment to one does not interfere with the outcome of another. In their foundational work, Hudgens and Halloran (2008) (following Sobel, 2006) developed an estimator that removes the assumption of no-interference between individuals by grouping the observations into fully connected subgraphs, that are not connected to each other. This subgraphs are also called clusters. In this strand of the literature the units of observation inside each cluster are connected to all the other ones inside the same cluster (and therefore interfere with each other), but not connected to those in different clusters (preventing the interference to spill over onto other clusters); this structure is usually referred to as *partial* or *clustered interference*. Cluster-randomised control trials are RCTs where the randomisation occurs at the cluster or subgraph level, not at the individual level.

The methodology developed by Hudgens and Halloran (also used in Halloran and Hudgens, 2012, 2016; Saul et al., 2017) allows for causal identification of the direct and spillover (indirect) effects when the assignment to treatment follows a two-stage randomisation procedure: first at the cluster level, and then at the individual level inside of each group. Tchetgen Tchetgen and VanderWeele (2012) extended these results into observational studies by using an inverse-probability-weighting estimator that corrects the lack of randomisation in the application of treatment at both levels. Building on these, Perez-Heydrich et al. (2014) and Papadogeorgou et al. (2019) calculate cluster-level direct and indirect causal treatment effects by considering many clusters comprised of few units of observation, where every one interacts with the others in the same cluster. Both Perez-Heydrich et al. and Papadogeorgou et al. estimate the average potential outcomes using inverse probability weights to account for differences between individuals assigned to treatment and control groups.

There are other approaches that address the presence of interference performing inference about estimands of interest. Athey et al. (2018) study how to calculate exact p-values for several null hypotheses where all units in the sample belong to a single connected network. Aronow and Samii (2017) present an inverse probability weighted estimator for the average unit-level causal effects from a randomised experiment with arbitrary but known interference. In their research, Aronow and Samii introduce the concepts of *treatment mapping* and *exposure mapping* to suggest a difference in who get assigned to treatment, and who actually gets exposed via the treatment that spills over. While treatment mapping refers to the way the experiment designers expect the intervention to be distributed through the population, the exposure mapping aims at understanding how it was actually distributed by taking into consideration the interference. One possible way of understanding this approach is that assuming no interference implies the treatment mapping and the exposure mapping are identical.

Ogburn and VanderWeele (2014) show that the kind of interference under consideration matters in terms of the assumptions needed for identification; furthermore, Ogburn (2017) and Lee and Ogburn (2020) suggest that the problems caused by statistical interference can be more widespread than initially assumed, since many studies do not meet the assumptions required by commonly used estimators. The most notable example is Ogburn et al. (2020), which challenges the long established result that obesity is socially contagious (see Christakis and Fowler (2007)). The interest in this particular area of research is growing, and the tools available for researchers looking into causal inference in the presence of interference are becoming more advanced. Further examples of this area of research can be found in Toulis and Kao (2013); Kang and Imbens (2016); Paluck et al. (2016); Athey and Imbens (2017); Jagadeesan et al. (2020).

2.9 Summary

In this chapter I presented an overview of the statistical concepts used in the rest of the dissertation divided into three major themes: social network analysis (an introduction and notation, the exponential random graph model and how to estimate it using maximum likelihood estimation as well as using Bayesian inference), relational event models (including randomisation inference for hypothesis testing), and causal inference (basic definitions and ideas, the potential outcomes framework, propensity scores, and causal inference in the presence of interference). These three themes correspond to the methodologies presented in three following chapters.

Chapter 3

Causal estimation of spillover effects in a social network setting: attempting to increase confidence in positive sexual health attitudes

3.1 Background on causal inference in the presence of interference

Small-scale public health interventions can be an important tool to investigate how to improve the lives of individuals in society. Given that many habits are formed during high school - adolescence is a period of increased openness (Kirby et al., 2007) - school interventions have been used to modify behaviour from an early age. Research has shown that although teachers or mentors are better at imparting factual information, behaviour and social norms of students can be more easily modified when guided by their peers (Harden et al., 2001). This suggests that peer-led interventions can be considered an alternative to more traditional, adult-led methods when we aim to influence adolescent behaviour.

The University of Glasgow's Social and Public Health Sciences Unit carried out a feasibility study for an intervention to improve sexual and reproductive health education in high-school students (Forsyth et al., 2018), known as Sexually Transmitted infections And Sexual Health (STASH). Aiming to augment school based education, the intervention intended to connect nominated students, trained as *peer supporters*, with other students to capitalise on the existing social relations in the school. This intervention followed steps previously proven to reduce the uptake of smoking among young people in the A Stop Smoking in Schools Trial (ASSIST) intervention. In that study, a cluster-randomised controlled trial with a similar peer-driven approach found that smoking was reduced over a two-year period (Campbell et al., 2008).

A key element to these interventions is the set of connections between students. I will refer to this as the *social network* of a group of individuals. Both STASH and ASSIST assume that units of observation (individual adolescents) are not independent from each other, but rather that their behaviour is guided by a structure of dependence, likely to be correlated with the network of connections between them (Rogers, 2002; Ogburn, 2017). By design, intervened upon individuals get in touch with their close connections and try to modify their behaviour. This presents a challenge to traditional causal inference research, which assumes that treatment given to one unit of observation does not affect the outcome of another unit. Here I am not only interested in correctly identifying the direct effect of the treatment. I am also interested in the effect of the treatment for those individuals that were not treated but who were exposed through their connections (these are called spillover effects), so I need to know how people are connected to each other.

The STASH feasibility study aimed to assess the recruitment and retention of peer supporters, fidelity and reach of the delivery of the intervention by trainers and peer supporters, as well as refine the theory behind the intervention. My aim is to estimate the causal direct effect of being a peer supporter, and investigate the effect of the intervention spilling over to other individuals in the sample, i.e. the causal indirect effect of the intervention. The intervention produced data with information on the individual characteristics of the students, the connections between them, and their treatment status. I will use that as a case-study for the regression approach proposed by Forastiere et al. (2020) that uses generalised propensity scores for determining causal direct and spillover effects of an intervention.

The main analysis carried out adopts the approach of Forastiere et al. (2020) (hereafter referred to as FAM for Forastiere, Airoldi, Mealli, the authors of the paper), as it was developed in a context similar to ours, where (fully-connected) clustered interference does not hold as not all individuals within a cluster are socially connected (for more on clustered interference see Hudgens and Halloran, 2008). As I detail in the next section, the approach relies on considering treatment as bivariate: direct treatment and indirect exposure to this treatment through social contacts, with generalised propensity scores (Hirano and Imbens, 2005) used to model each component of this bivariate treatment separately and then included these as covariates in an outcome regression model.

The direct effect estimates what happens, on average, if individuals received the treatment versus if they did not, in counterfactual terms. This is, estimating the average treatment effect. Similarly, for the indirect effect, we can estimate the average spillover effect for both units that were treated as well as those untreated.

The relevant literature for causal inference in the presence of interference can be found in Section 2.8, and in particular a practical implementation of FAM's methodology can be found in Del Prete et al. (2020). Section 3.2 presents a description of the methodology used by FAM to determine the causal direct and spillover effect of an intervention. I introduce an alternative way of estimating the direct and spillover effects by way of cubic splines in the outcome model. My proposal uses flexible modelling as an alternative to the reliance of the FAM estimators on researchers' *a priori* understanding of the functional form of the outcome model. As I shall detail in the methodology below (Section 3.2.2), one of the main requirements for the FAM methodology to produce unbiased estimators is full knowledge of the form of an outcome model, as well as two different treatment models. The first treatment model relates to the direct treatment, while the second one relates to the way individuals are exposed to the treatment indirectly via their social connections with the peer supporters. Peer supporters can also experience spillover from their treated connections.

In Section 3.3, I investigate the sensitivity of the FAM approach to violations of the assumption of full information of the outcome model, and present my alternative modelling strategy. I expand on the work of the authors by estimating the direct and spillover effects using directed networks, as opposed to undirected ones, and show that the suggested subclassification approach is limited by the sample size. Section 3.4 presents a description of the STASH intervention, an overview of the data, and the methodologies that will be used to estimate STASH's main and spillover effects. In Section 3.5, I discuss my findings.

3.2 Propensity-score regression for data with interference

For this analysis, I will consider the potential outcomes framework. An important assumption in traditional causal inference research is that the treatment applied to one unit of observation only affects its own outcome, and does not affect the outcome of other units in the sample. The literature refers to this as the no-interference, or no spillover assumption, and it is a core component of an assumption made by many causal approaches: the Stable Unit Treatment Value Assumption, or SUTVA (Cox, 1958 and Rubin, 1980). To recap, the Stable Unit Treatment Value Assumption has two parts. The first assumes that there are not multiple versions of treatment (also known as consistency), such that the potential and realised outcomes under the same treatment coincide. Forastiere et al. (2020) frame it as "the mechanism used to assign the treatments does not matter and assigning the treatments in a different way does not constitute a different treatment". The second refers to the lack of interference between units of observation. Formally, this is that the outcome of one individual which could depend on the treatment to everyone in the sample, $W_i(\mathbf{Z})$, only depends on the treatment applied to that individual, $W_i(Z_i)$.

Under many scenarios, when the assumption of no interference is valid, it is possible to obtain unbiased estimators for treatment effects. However, in the case of inherently interconnected data this assumption is harder to justify, suggesting the potential for bias and incorrect inference. Consider the following example: if (by design) treated individuals expose control individuals to the treatment (like in the STASH study, where treatment consists of educational material provided to peer supporters who are then requested share that to their peers), we would like to know what the effect of the intervention was on the units in the first group, but it is also of interest how the indirect exposure to the treatment affected the outcome of the units in the second group. Assuming that the outcomes of this second group of students were not affected by the treatment would result in an underestimation (or overestimation, depending on the specifics of the intervention) of the true effect.

In this regard, interference presents a challenge for statistical inference because the potential outcomes for each individual depend on the individual treatment, as well as the (spilled over) treatment received by all the units, which violates the second component of SUTVA. Forastiere et al. (2020) developed a method to determine the direct and spillover effects from an intervention using observational data, considering the treatment as a bivariate vector in which the balance of each covariate for each component of the vector is achieved via (generalised) propensity score methods (following Hirano and Imbens, 2005).

3.2.1 Potential outcomes framework under interference

Extending the definition of potential outcomes originally laid out by Rubin (1974), Forastiere et al. (2020) suggest a modified version of SUTVA called Stable Unit Treatment on Neighbourhood Value Assumption (SUTNVA). Similar to previous work by van der Laan (2014), they assume that the propagation of treatment only occurs between immediate connections (i.e., a person may be influenced by their friends, but not by friends of their friends). The causal estimands proposed by the authors are average comparisons of the potential outcomes under different combinations of the treatment and the interference.

SUTNVA and the rest of our analysis rely the following notation and definitions introduced in Section 2.2, recalled here:

- A **network** is a pair $(Y, \mathbf{X})$, composed of a set n nodes, and a set of edges, or connections between those nodes, represented by Y, and $\mathbf{X}$ represents a matrix with n rows and k columns, representing the values of the k attributes for each of the n nodes in the network. For a given pair of nodes i and j, the pair (i, j) represents a connection between i and j. Networks can be "directed", which means that $(i, j) \in \mathcal{Y} \not\Rightarrow (j, i) \in \mathcal{Y}$. In this case, i *sends* a connection to j (j has an *incoming* connection from i). For "undirected" networks, $(i, j) \in \mathcal{Y} \Leftrightarrow (j, i) \in \mathcal{Y}$.

- $\mathcal{N}$ is the set of nodes in the network.

- $\mathcal{N}$ can be partitioned as $(i, \mathcal{N}_i, \mathcal{N}_{-i},)$, where $\mathcal{N}_i$ refers to all the nodes in $\mathcal{N}$ that are connected by links incoming to i, and $\mathcal{N}_{-i}$ are all the nodes that are not connected by incoming links to i. In the case of undirected networks, $\mathcal{N}_i$ is referred to as the **neighbourhood** of unit i. In directed networks, we use *incoming* neighbourhood when considering the incoming edges to i, or *outgoing* neighbourhood for the outgoing edges from i.

- $Z_i \in \{0, 1\}$ is a binary random variable representing the treatment of unit i, 0 if not treated and 1 if treated. $\mathbf{Z}$ is the vector that contains the treatment assignments for all nodes in $\mathcal{N}$.

- Similarly, W_i is a random variable representing the observed outcome of unit i. $\mathbf{W}$ is the vector that contains the outcomes for all nodes in $\mathcal{N}$[1].

- $\mathbf{X}_i$ is the vector of covariates for node i. $\mathbf{X}_i^{ind}$ refers to a vector of individual covariates for node i, and $\mathbf{X}_i^{net}$ refers to a vector of network-related covariates for observation unit i. $\mathbf{X}^{net}$ can, for example, be the average value of $\mathbf{X}_{\mathcal{N}_i}^{ind}$ for the individuals in the neighbourhood of i, or the number of incoming connections of i, $||\mathcal{N}_i||$.

- Neighbourhood treatment: this is the treatment that spills over onto unit i from the units in the sample belonging to the incoming neighbourhood of i. For clarity, this is a measure of treatment and not an estimand of an effect.

Under interference, the observed outcome for unit i, W_i, may be a function of the entire treatment assignment vector $\mathbf{Z}$, or $W_i = W_i(\mathbf{Z})$. The first component of SUTVA, that there is but a

[1] Most of the causal inference literature refers to the outcome of a unit with the variable Y, however, for consistency with the rest of the dissertation I change this to W.

single version of treatment, implies that assigning treatments in different ways does not constitute a different treatment. SUTNVA does not modify this assumption. In other words, the mechanism of assigning a treatment to individuals does not change the outcome even in spite of the fact that the outcome explicitly depends on the entire vector of treatments. van der Laan (2014) and Forastiere et al. (2020) relaxed this somewhat such that the outcome of individual i depends not on the entire vector of treatments, but only on the treatments of the incoming neighbourhood of i, or explicitly only those nodes of the neighbourhood of i whose edges are directed *into* i.

To see how this assumption is defined in terms of network interference, we can divide our sample in three main parts: a single unit, i, its incoming neighbourhood, $\mathcal{N}_i$, and all the units that are not in its incoming neighbourhood $\mathcal{N}_{-i}$. The "no-interference" assumption implies that $W_i(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}) = W_i(Z_i, \mathbf{Z'}_{\mathcal{N}_i}, \mathbf{Z'}_{\mathcal{N}_{-i}})$ for all different $\mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}, \mathbf{Z'}_{\mathcal{N}_i}, \mathbf{Z'}_{\mathcal{N}_{-i}}$. In other words, *nobody's* exposure but i's matters.

However, in our case, we are interested in how i is exposed to the treatment assigned to the units in its incoming neighbourhood, $\mathcal{N}_i$.

Forastiere et al. (2020) propose a modified version of the no-interference assumption: Consider

$$g_i : \{0, 1\}^{N_i} \to \mathcal{G}_i,$$

a function that takes the vector of all possible treatments for a group of individuals in an incoming neighbourhood, $\mathcal{N}_i$, $\{0, 1\}^{N_i}$, with N_i being the number of nodes in $\mathcal{N}_i$, and produces an aggregate version of that treatment applied to that group of units. $\mathcal{G}_i$ is the domain of g_i and ultimately depends on the definition of the function and what we mean by "aggregation". For example, in one case, g_i could the *number* of treated peers in $\mathcal{N}_i$, $\mathcal{G}_i = \{0, \cdots, N_i\}$, and in another, g_i could be the *proportion* of treated peers in $\mathcal{N}_i$, which makes $\mathcal{G}_i = [0, 1]$.

We can then say that for all $\mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}'_{\mathcal{N}_i}$ (different treatment allocations of the incoming neighbourhood of i) and for all $\mathbf{Z}_{\mathcal{N}_{-i}}, \mathbf{Z}'_{\mathcal{N}_{-i}}$ (different treatment allocations for the units outside of the neighbourhood of i), if the aggregation of the treatment allocation to units in $\mathcal{N}_i$ (that is, $\mathbf{Z}_{\mathcal{N}_i}$) is the same as the aggregation of a different allocation of treatment to the same units ($\mathbf{Z}'_{\mathcal{N}_i}$), in other words, that $g_i(\mathbf{Z}_{\mathcal{N}_i}) = g_i(\mathbf{Z}'_{\mathcal{N}_i})$, then unit i's outcome is the same under both treatment allocations,

$$W_i(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}) = W_i(Z_i, \mathbf{Z'}_{\mathcal{N}_i}, \mathbf{Z'}_{\mathcal{N}_{-i}}),$$

regardless of the values of $\mathbf{Z}_{\mathcal{N}_{-i}}$ and $\mathbf{Z}'_{\mathcal{N}_{-i}}$

Under the proposed assumptions, FAM only required that the potential outcomes be indexed by the individual treatment Z_i and the neighbourhood treatment of unit i, $G_i = g_i(\mathbf{Z}_{\mathcal{N}_i})$. $\mathcal{G}$ is the space of all possible values of G_i. Because of this, the potential outcome for unit i, $W_i(z, g)$, can only be calculated for the subset of nodes where G_i can be take value g, defined as V_g by the authors. The cardinality of V_g is expressed as v_g. In other words, if g_i is a function that aggregates neighbourhood treatment as the share of treated peers, and unit i only has two peers, it will not belong to $V_{0.25}$, since there is no way for that unit to be exposed by a quarter of its peers, and so the potential outcome for that unit with $g = 0.25$ is not defined.

To clarify, I use the same example provided by the authors: "in the case where G_i is the number of treated neighbours, V_g is the set of nodes with degree $N_i \geq g$." Note this is just an example of the how V_g can be expressed based on the definition of G_i. Different G_i's lead to different definitions of V_g.

Regarding the individual direct effect of treatment for a particular level of interference g, the causal estimand can be expressed as

$$\tau(g) = E\left[W_i(Z_i = 1, G_i = g) - W_i(Z_i = 0, G_i = g) | i \in V_g\right],$$

where it must be noted that this effect is conditional on a given value g. This is, the effect of individual treatment, $(Z_i = 1)$ vs $(Z_i = 0)$, when the neighbourhood treatment level is set to g.

The overall, or marginal, direct effect τ is calculated by averaging the individual treatment over the probability distribution of the neighbourhood treatment:

$$\tau = \sum_{g \in \mathcal{G}} \tau(g) P(G_i = g).$$

Regarding the neighbourhood effect, the authors define this as the spillover effect of having a neighbourhood treatment effect set to g versus 0, when the unit is under individual treatment z:

$$\delta(g, z) = E\left[W_i(Z_i = z, G_i = g) - W_i(Z_i = z, G_i = 0) | i \in V_g\right].$$

Similarly, the overall spillover effect, marginalised over G, is defined as

$$\Delta(z) = \sum_{g \in \mathcal{G}} \delta(g, z) P(G_i = g), \text{ for } z = 0, 1.$$

The direct and spillover effects are based on the comparison between the marginal (over the covariates) mean of two different potential outcomes, and represent the values I want to estimate. In the case of the direct effect, $\tau(g)$ compares, for a given neighbourhood treatment, the effect of being assigned to treatment versus not being assigned to treatment. In the case for the neighbourhood treatment effect, $\delta(z, g)$ compares, for a given level of individual treatment, the effect of a set level of neighbourhood treatment, versus receiving no neighbourhood treatment at all. This marginal (again over the covariates) mean of the potential outcome for a subset V_g is defined as

$$\mu(z, g) = E[W_i(z, g) | i \in V_g], \ \forall z \in 0, 1, g \in \mathcal{G}.$$

The innovation proposed by Forastiere et al. is that $\mu(z, g)$ can be seen as an average dose-response function depending on the dose of a bivariate treatment, where one component of the treatment vector captures the direct effect and the other the indirect or spillover effect, with marginalisation occurring over the space of the covariate vector, X. I detail the estimation procedure from Forastiere et al. (2020) in the next chapter.

3.2.2 (Generalised) propensity score regression for direct and spillover treatment effects

The method proposed by Forastiere et al. (2020) considers two different propensity scores: one that determines the individual treatment, and one that determines the neighbourhood treatment. Consider the joint distribution of the bivariate treatment vector: $\varphi(z, g|x) = P(Z_i = z, G_i = g|X_i = x)$, the probability for unit i of being assigned to individual treatment z and exposed to neighbourhood treatment g, given the observed characteristics x.

The authors establish that the joint propensity score for these two kinds of treatment is: i) a balancing score, which means that if a set of units differ in terms of their direct treatment z or their neighbourhood treatment g but have the same value of $\varphi_i(z, g|x)$, the distribution of the covariates x is the same; ii) conditionally unconfounding of Z_i and G_i, which means that the outcome W_i is conditionally independent of the individual and the neighbourhood treatments given $\varphi(z, g|x)$; and iii) factorisable as

$$\begin{aligned}\varphi(z, g|x) &= P(Z_i = z, G_i = g|X_i = x)\\ &= P(G_i = g|Z_i = z, X_i^g = x^g)P(Z_i = z|X_i^z = x^z)\end{aligned}$$

where X^z refers to the group of variables that estimate the *individual* propensity score, $P(Z_i = z|X_i^z = x^z)$, and X^g is the set of variables that estimate the *neighbourhood* propensity score, $P(G_i = g|Z_i = z, X_i^g = x^g)$. These two sets of variables do not need to be the same. The probability of having neighbourhood treatment at level g conditional on a specific value z of the individual treatment and on X^g, $P(G_i = g|Z_i = z, X_i^g = x^g)$, is denoted by $\lambda_i(g|z_i, x_i^g)$, and is referred to as the neighbourhood propensity score. The probability of having individual treatment at level z conditional on X^z, $P(Z_i = z|X_i^z = x^z)$ is represented by $\phi_i(z|x_i^z)$ and is referred to as the individual propensity score.

The estimation strategy proposed by Forastiere et al. (2020) proceeds as follows:

1. Estimate the individual propensity score, $\phi(z|x^z)$, using a logistic regression for the treatment conditional on X^z.
2. Predict the propensity to be selected into treatment, $\hat{\phi}_i(z|x_i^z)$ for each unit.
3. Subclassify the data based on the predicted individual propensity score, $\hat{\phi}(z|x^z)$ into J parts.
4. For each subclass j, estimate $\mu_j(z, g) = E[W_i(z, g)|i \in B_j^g]$, where B_j^g is the collection of units in subclass j, and in V_g. These values will then be used to calculate the overall $\mu(z, g)$, i.e. dose-response function. To get each $\mu_j(z, g)$ we:
    (a) Estimate parameters for the neighbourhood treatment, assuming a specific distribution for the neighbourhood effect. The original functional form proposed was:

$$logit(\lambda^{[j]}(g|z, X^g)) = \gamma_0^{[j]} + \gamma_Z^{[j]} Z_i + \gamma_{\mathbf{X}_g}^{[j]\prime} \mathbf{X}_i^g,$$

but others are also allowed. Predict individual values of $\lambda(g, z|X^g)$ inside of each subclass j, and refer to these as $\hat{\lambda}_i^{[j]}$, for individual i in subclass j.

(b) Use the observed data (W_i, Z_i, G_i, X_i) and the estimated $\hat{\lambda}_i^{[j]}$ to estimate the parameters of a model for the potential outcome $W_i(z, g)$ as a function of treatment, of the individual and neighbourhood propensity score, and also of covariates. FAM proposed a parametric outcome model that depends on the individual treatment, the neighbourhood treatment, and the estimated neighbourhood propensity score:

$$W_i(z, g) = f(z_i, g_i, \hat{\lambda}^{[j]}(z_i, g_i, X_i^g)). \tag{3.1}$$

The actual functional form of f depends on the actual outcome we want to analyse. The estimation of $\hat{\lambda}_i^{[j]}$ can present complications due to the small size of the strata in studies such as STASH.

(c) For a particular level of the joint treatment $(Z_i = z, G_i = g)$, for each unit in the subclass we predict the neighbourhood propensity score evaluated at that level of the treatment, i.e., $\hat{\lambda}_i(g|z, X_i)$, and use it to predict the potential outcome $W_i(z, g)$ using the parameters estimated in step (b). This is the most important step of the entire estimation, since this is where the potential outcomes are calculated.

(d) Estimate the subclass-specific dose-response values, $\hat{\mu}_j(z, g, V_g)$ by averaging the individual potential outcomes for every combination of individual and neighbourhood treatment:

$$\hat{\mu}_j(z, g, V_g) = \frac{\sum_{i \in B_j^g} \hat{W}_i(z, g)}{|B_j^g|}.$$

5. With all the $\mu_j(z, g)$, calculate the average dose-response function as the weighted average of the subclass-specific dose-response values, where the weights are calculated as the proportion of individuals in each class:

$$\hat{\mu}(z, g, V_g) = \sum_{j=1}^{J} \hat{\mu}_j(z, g, V_g) \left(\frac{|B_j^g|}{v_g} \right).$$

6. With the dose-response average potential outcome, estimate the direct and spillover effects of the intervention.

The subclassification mentioned in step three can be done in different ways, always looking for balance (where the average value of the observable variables, x^z, is the same for all subclasses (Cunningham, 2021)) between the subclasses. Having a large number of subclasses makes balance harder to achieve when samples are small because of the lower number of observations in each subclass. FAM reports using two subclasses for their estimation.

A clear challenge presented by this methodology is that it requires correct specification both in the individual treatment model. Another challenge is requiring the correct specification of the outcome model. I am going to study the behaviour of the estimator proposed by Forastiere et al. (2020) considering different model specifications. Additionally, the simulations presented by those authors have a relatively large sample size. I replicate their simulation using diverse sample sizes and use directed rather than undirected networks.

3.2.3 Flexible regression using cubic splines

As an alternative model specification, I consider a flexible regression on the outcome model, with cubic splines on the estimated individual and neighbourhood propensity scores in addition to terms for the treatment, spillover variables and their interaction. This means including a cubic polynomial for both propensity scores, and introducing a knot at the median of the variable. The flexibility provided by this modelling specification reduces the need to know the exact functional form of the model we are estimating, while capturing the complexity of the interaction between the outcome and the propensity scores.

Using splines in the outcome model should allow for greater flexibility in the estimation of the direct and spillover effects, without having complete information of the data generating process. I am interested in including the variables that both affect the outcome and the way treatment is distributed throughout the sample both directly and by means of the spillover. A flexible regression approach includes these variables in an agnostic way, requiring less information about the outcome model from the researcher who intends to use this methodology. My investigation led to using two comparable flexible regressions: one that considers the subclasses, as determined by the authors

(Section 3.2.2, step 3), and one that does not. This subsection presents the details of my proposed approach.

For clarity, I first explain our approach without subclasses, and then the one with subclasses:

Approach without subclasses:

1. Estimate the individual propensity score, $\phi(z|x^z)$, using a logistic regression for the treatment conditional on X^z (same as step 1 in the FAM approach).

2. Predict the propensity to be selected into treatment, $\hat{\phi}(z|x^z)$ for each unit (same as step 2 in the FAM approach).

3. Estimate the potential outcomes for W_i when $Z_i = 0$ and when $Z_i = 1$. To do this, estimate parameters for the neighbourhood treatment model, according to the proposed functional form:
$$logit(\lambda(g|z, X^g)) = \gamma_0 + \gamma_Z Z_i + \gamma_{x_g} X_i^g.$$

4. Use the following flexible outcome model for the calculation of the potential outcomes:
$$W_i(Z_i, G_i) \sim \beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i G_i + \beta_{\hat{\phi}} \text{spline}(\hat{\phi}_i) + \beta_{\hat{\lambda}} \text{spline}(\hat{\lambda}_i),$$
where $\hat{\phi}$ is the predicted propensity score for the treatment model, and $\hat{\lambda}$ is the predicted propensity score for the neighbourhood model when $Z = 0$ and $Z = 1$. Similarly to the way the dose-response function was calculated in Subsection 3.2.2, I calculate $W_i(Z_i = z, G_i = g)$ for every value of z and g available.

5. Predict individual values of $\lambda(g, Z_i = z|X^g)$, and refer to them as $\hat{\lambda}_i(Z_i = z)$. This is then used to estimate the potential outcome $\hat{W}_i(Z_i = z, G_i = g)$, following the same approach described in the estimation strategy above in steps 4.c and 4.d.

6. With the dose-response average potential outcome, estimate the direct and spillover effects of the intervention.

Approach with subclasses: This approach works very similarly to the one without subclasses, with the particular distinction that we perform the estimation of the potential outcomes (therefore estimate the direct and spillover effects) inside of each subclass. The final estimate is the weighted average of the direct and spillover effects, where the weights are given by the number of individuals in each subclass.

1. Estimate the individual propensity score, $\phi(z|x^z)$, using a logistic regression for the treatment conditional on X^z.

2. Predict the propensity to be selected into treatment, $\hat{\phi}(z|x^z)$ for each unit.

3. Subclassify the data based on the predicted individual propensity score, $\hat{\phi}(z|x^z)$ into J parts.

4. For the individuals whose value of $\hat{\phi}(z|x^z)$ lies in subclass j, and for every subclass:

   (a) Estimate the potential outcomes for $W_i^{[j]}$ when $Z_i = 0$ and when $Z_i = 1$. To do this, we first need to estimate parameters for the neighbourhood treatment model, according to the proposed functional form:

   $$logit(\lambda^{[j]}(g|z, X^g)) = \gamma_0^{[j]} + \gamma_Z^{[j]} Z_i + \gamma_{x_g}^{[j]\prime} X_i^g,$$

   to then include it the following flexible outcome model:

   $$W_i^{[j]}(Z_i, G_i) \sim \beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i G_i + \beta_{\hat{\phi}} \text{spline}(\hat{\phi}_i) + \beta_{\hat{\lambda}} \text{spline}(\hat{\lambda}_i^{[j]}),$$

   where $\hat{\phi}$ is the predicted propensity score for the treatment model, and $\hat{\lambda}^{[j]}$ is the predicted propensity score for the neighbourhood model, calculated inside of the j-th subclass. Similarly to the way the dose-response function was calculated in Subsection 3.2.2, we calculate $W_i^{[j]}(Z_i = z, G_i = g)$ for every value of z and g available.

   (b) This is, we predict individual values of $\lambda^{[j]}(g, Z_i = z|X^g)$ inside of each subclass j, and refer to them as $\hat{\lambda}_i^{[j]}(Z_i = z)$. This is then used to estimate the potential outcome $\hat{W}_i^{[j]}(Z_i = z, G_i = g)$.

   (c) To calculate the potential outcomes, follow the same approach described in the estimation strategy above in steps 4.c and 4.d. I estimate the subclass-specific dose-response values, $\hat{\mu}_j(z, g, V_g)$ by averaging the individual potential outcomes for every combination of individual and neighbourhood treatment:

$$\hat{\mu}_j(z, g, V_g) = \frac{\sum_{i \in B_j^g} \hat{W}_i(z, g)}{|B_j^g|}.$$

5. With all the $\mu_j(z, g)$, we calculate the average dose-response function as the weighted average of the subclass-specific dose-response values, where the weights are calculated as the proportion of individuals in each class:

$$\hat{\mu}(z, g, V_g) = \sum_{j=1}^{J} \hat{\mu}_j(z, g, V_g)\left(\frac{|B_j^g|}{v_g}\right).$$

6. With the dose-response average potential outcome I estimate the direct and spillover effects of the intervention.

3.3 Simulation

I present a simulation study to examine how the discussed estimation methodologies perform when calculating the direct and spillover effects of an intervention and compare alternative approaches, while also considering model misspecification. In this section I present the data generating process and the fitting of the models.

3.3.1 Data generating process

Our simulation is designed similarly to that of Forastiere et al. (2020) and is used compare the results when the drawn sample is large to when the sample size is more in line with the number of observations in the STASH trial. Importantly, I also test the robustness of the method to misspecifications in the models used by the data analyst. The unit of observation is going to be a student in a school. Schools work like clusters (in the Hudgens and Halloran, 2008 sense) in that the transmission of treatment between schools should not be allowed, meaning that they are distinct networks with no connections between them. I do this to try to replicate the way in which the data for the STASH intervention was collected. I now explain the simulation and the estimation procedure.

The number of schools/clusters is going to determine the ultimate sample size. Every school is generated with 100 students, and each student has two independent variables $x_1^{ind} \sim \text{Bernoulli}(1, 0.3)$ and $x_2^{ind} \sim \text{Poisson}(\lambda = 2)$. Within each school, we sample a friendship network between the students using the `network` package (Butts et al., 2021). The formula used for the generation of all of these networks was

$$\text{network} \sim \text{edges} + \text{mean degree} + \text{gwesp}(0.5),$$

with coefficients (-4, -7, -0.5). The set of network terms as well as their respective coefficients were chosen because they produce networks compatible with the ones from the STASH study (see Table A.1 and Figure A-1 in the Appendix for this chapter). For this simulation, I consider a "treated individual" to be those that receive the treatment directly, and those individuals who are in the neighbourhood of the treated individuals are considered "exposed individuals".

Following the restriction on the STASH friendship questionnaire, we cap the maximum number of friends at 5. With the information on everyone's friends, following FAM, we calculate the average values of x_1^{ind} and x_2^{ind} across each individual's incoming connections to create x_1^{net} and x_2^{net}, respectively. I use the average to generate x_1^{net} and x_2^{net}, following Forastiere et al. (2020).

The variables that correspond to individual information are noted as x_1^{ind} and x_2^{ind}, and I refer to them as X^{ind}. The variables that correspond to the network information for each individual are noted as x_1^{net} and x_2^{net}, and referred to as X^{net}. The number of incoming friendships each individual has is referred to as that individual's in-degree. Reciprocally, the number of outgoing friendships each individual has is referred to as that unit's out-degree.

Individual treatment, z_i, is generated as a Bernoulli(μ_{z_i}), where μ_{z_i} is modelled as follows:

$$logit(\mu_{z_i}) = -12 + 1.5x_1^{ind}{}_i + 3x_2^{ind}{}_i - 0.5x_1^{net} + x_2^{net}.$$

x_{1i}^{ind} is the value of x_1^{ind} for individual i. The same notation is used for all other variables. Under this scenario, interference is determined by the proportion of peers that are treated that each individual has. I calculate this by dividing the number of treated *incoming* treated peers by the number of incoming connections for each individual:

$$g_i(X_{N_i}) = G_i = \frac{\text{treated peers}_i}{\text{in-degree}_i}.$$

Following the definition introduced in Section 3.2, when I mention the *neighbourhood of student i*, it is in reference to the incoming neighbourhood of i, as it was defined in Section 3.2.1

The outcome, a variable that increases with treatment and exposure to treatment, is determined by several influences including individual connections and the connections' treatment status, which is in turn determined by their covariates

$$W_i(z, g)|X^{ind}, X^{net} \sim \mathrm{N}(\mu_{W_i}(Z, G, X^{ind}, X^{net}), 1)$$

where

$$\begin{aligned} \mu_{W_i}(Z, G, X^{ind}, X^{net}) = \\ & 33 + x^n_{i1} + 7x^n_{i2} - 10\mathbb{1}(\phi_i(Z_i|X^{ind}_i, X^{net}_i) \geq 0.7) \\ & + 10Z_i + \delta G_i - 10\hat{\lambda}_i(G_i|z_i, X^{ind}_i, X^{net}_i) + 5G_i\mathbb{1}(\phi_i(Z_i|X^{ind}_i, X^{net}_i) \geq 0.7) + 3Z_iG_i. \end{aligned} \tag{3.2}$$

There are several things to be highlighted from Equation 3.2. The first is that in these simulations, $\hat{\lambda}(G|Z, X^{ind}, X^{net})$ is not the data-generating "propensity score" for G, but rather a prediction of G given Z, X^{ind}, and X^{net}, fit after G is observed in the data-generating process.

The second is that $-\delta G_i + 5G_i\mathbb{1}(\phi_i(X^{ind}, X^{net}) \geq 0.7) + 3Z_iG_i$ represents the spillover effect to unit i from its peers. Additionally, the inclusion of the indicator function taking a value of 1 when the predicted propensity score of the individual treatment exceeds 0.7, creates a very precise break exploited by the subclass methodology. I propose a second, alternative data generating model without a step function and compare the results for all estimation procedures and both data generating processes in Equation 3.3.

$$\mu_{W_i}(Z, G, X^{ind}, X^{net}) =$$

$$33 + x^n_{i1} + 7x^n_{i2} - 10(\phi_i(Z_i|X^{ind}_i, X^{net}_i))^3$$
$$+ 10Z_i + \delta G_i - 10\lambda_i(G_i|z_i, X^{ind}_i, X^{net}_i) + 5G_i(\phi_i(Z_i|X^{ind}_i, X^{net}_i))^3 + 3Z_iG_i. \tag{3.3}$$

According to this model, the true direct effect τ^* and the true spillover effect $\delta^*(z)$, as functions of Z and G, are the terms in the outcome model where the direct treatment and the spilt-over treatment are present. The direct effect for a particular level of interference, g is

$$\tau^*(g) = -10 + 3g. \tag{3.4}$$

The overall true direct effect, across all values of G, is calculated as an expected value over G,

$$\tau^* = \sum_{g \in G} \tau(g)P(G_i = g).$$

In the specific case of this data generating process, this value is $\tau^* = -10 + 3E[G_i]$.

The spillover effect given a particular level of interference and treatment value, $\delta^*(z, g)$, is

$$\delta^*(z, g) = -\delta G_i - 10\lambda_i(G_i, Z_i, X^{ind}_i, X^{net}_i) + 5G_i\mathbb{1}(\phi_i(X^{ind}_i, X^{net}_i) \geq 0.7) + 3Z_iG_i, \tag{3.5}$$

for the original data generating process, and

$$\delta^*(z, g) = -\delta G_i - 10\lambda_i(G_i, Z_i, X^{ind}_i, X^{net}_i) + 5G_i(\phi_i(Z_i|X^{ind}_i, X^{net}_i))^3 + 3Z_iG_i, \tag{3.6}$$

for the new generation process.

In both Equation 3.5 and 3.6, δ is the impact of the share of treated peers each student has. $\delta^*(z, g)$ is calculated as the contrast between the potential outcome considering $G = g$ and $G = 0$. The general spillover effect across all values of g for a particular value of z, $\Delta^*(z)$, is calculated as an expected value like with the direct effect.

$$\Delta^*(z) = \sum_{g \in G} \delta^*(g, z) P(G_i = g)$$

For the specific case of these two data generating processes, this is:

$$\Delta^*(z) = \delta E[G_i] - 10E[\lambda(G_i|Z_i, X_i^{ind}, X_i^{net})] + 5E[G_i]E[\mathbb{1}(\phi_i(Z_i|X_i^{ind}, X_i^{net}) \geq 0.7)] + 3Z_i E[G_i],$$

and

$$\Delta^*(z) = \delta E[G_i] - 10E[\lambda(G_i|Z_i, X_i^{ind}, X_i^{net})] + 5E[G_i]E[(\phi_i(Z_i|X_i^{ind}, X_i^{net}))^3] + 3Z_i E[G_i],$$

respectively.

Estimation

Several estimation procedures are used to try to recover the true values of the direct and spillover effect. I will consider three scenarios with methodology proposed by Forastiere et al. (2020): in the first, I specify the individual, the neighbourhood propensity score and the outcome model as the correct (generated) models; this scenario is referred to as "FAM (correct)". In the second scenario, I specify the correct individual and neighbourhood propensity score models, but under-specify the true outcome model by omitting x_1^n and x_2^n from Equation 3.1. I refer to this scenario as "FAM (incorrect outcome)".

The third scenario, which is referred to as "FAM (incorrect PS)", considers the correct outcome model, but allows for misspecification of the individual treatment model by omitting x_2^n. I do not expect this exclusion to produce assumption violations since we are including the variable in the outcome model, i.e. there is no omitted variable bias.

Following the methodology described in Subsection 3.2.2, I present the results using the three models described above. Given that I am calculating the outcome models inside of the subclasses in two of the three estimation methods, the estimated coefficients have a superscript indicating in which subclass they are being calculated, this is, $\beta_{x_2^n}^{[j]}$ is the coefficient for x_2^n in subclass j.

In addition to the estimate based on generalised propensity scores, I calculate the direct and spillover effects of the intervention using unadjusted and adjusted linear models. The details for all of these estimation methods is found in Table 3.1.

For the unadjusted and the adjusted models, I recover the direct effect as $\hat{\beta}_z + \hat{\beta}_{zg} z \hat{E}[G_i]$, and the spillover effect as $\hat{\beta}_g z \hat{E}[g_i]$ when $Z = 0$, and $(\hat{\beta}_g + \hat{\beta}_{zg}) z \hat{E}[G_i]$ when $Z = 1$. $\hat{E}[G_i]$ is the sample average. The neighbourhood propensity score model and the outcome model in the FAM estimates have a j superscript indicating that the models are fit inside of every subclass. In the flexible regression models, the j superscript is only present when I estimate it using subclasses, similar to the FAM methodology.

Table 3.1: Summary of estimation methods for calculation of direct and spillover effects. SC = Subclasses

Name	Outcome model ($E[W_i\|.] =$)	Individual PS ($logit(E[Z_i\|.]) =$)	Neighbourhood PS ($logit(E[G\|.])$)
Unadjusted	$\beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i G_i$	NA	NA
Adjusted	$\beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i G_i + \beta_{X^{ZG}} \mathbf{X}_i^{ZG}$	NA	NA
Flexible (correct)	$\beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i\ G_i +$ $\beta_{\hat{\phi}} z\text{spline}(\hat{\phi}_i) + \beta_{\hat{\lambda}} z\text{spline}(\hat{\lambda}_i)$	$\alpha_{x_1^{ind}} x_{i1}^{ind} + \alpha_{x_2^{ind}} x_{i2}^{ind} + \alpha_{x_1^{net}} x_{i1}^{net} + \alpha_{x_2^{net}} x_{i2}^{net}$	$\beta_Z Z_i + \beta_{in-degree} in-degree_i +$ $\gamma_{x_1^{ind}} x_{i1}^{ind} + \gamma_{x_2^{ind}} x_{i2}^{ind} + \gamma_{x_1^{net}} x_{i1}^{net} +$ $\gamma_{x_2^{net}} x_{i2}^{net}$
Flexible (incorrect PS)	$\beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i z G_i +$ $\beta_{\hat{\phi}} z\text{spline}(\hat{\phi}_i) + \beta_{\hat{\lambda}} z\text{spline}(\hat{\lambda}_i)$	$\alpha_{x_1^{ind}} x_{i1}^{ind} + \alpha_{x_2^{ind}} x_{i2}^{ind} + \alpha_{x_1^{net}} x_{i1}^{net}$	
Flexible (correct) - SC	$\beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i z G_i +$ $\beta_{\hat{\phi}} z\text{spline}(\hat{\phi}_i) + \beta_{\hat{\lambda}} z\text{spline}(\hat{\lambda}_i^{[j]})$	$\alpha_{x_1^{ind}} x_{i1}^{ind} + \alpha_{x_2^{ind}} x_{i2}^{ind} + \alpha_{x_1^{net}} x_{i1}^{net} + \alpha_{x_2^{net}} x_{i2}^{net}$	$\beta_Z^{[j]} Z_i + \beta_{in-degree}^{[j]} in-degree_i +$ $\gamma_{x_1^{ind}}^{[j]} x_{i1}^{ind} + \gamma_{x_2^{ind}}^{[j]} x_{i2}^{ind} + \gamma_{x_1^{net}}^{[j]} x_{i1}^{net} + \gamma_{x_2^{net}}^{[j]} x_{i2}^{net}$
Flexible (incorrect PS) - SC	$\beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i z G_i +$ $\beta_{\hat{\phi}} z\text{spline}(\hat{\phi}_i) + \beta_{\hat{\lambda}} z\text{spline}(\hat{\lambda}_i^{[j]})$	$\alpha_{x_1^{ind}} x_{i1}^{ind} + \alpha_{x_2^{ind}} x_{i2}^{ind} + \alpha_{x_1^{net}} x_{i1}^{net}$	
FAM (correct) - SC	$\beta_Z^{[j]} Z_i + \beta_G^{[j]} G_i + \beta_{ZG}^{[j]} Z_i G_i +$ $\beta_{x_1^{net}}^{[j]} x_{i1}^{net} + \beta_{x_2^{net}}^{[j]} x_{i2}^{net} + \beta_{\hat{\lambda}}^{[j]} \hat{\lambda}_i^{[j]}$	$\alpha_{x_1^{ind}} x_{i1}^{ind} + \alpha_{x_2^{ind}} x_{i2}^{ind} + \alpha_{x_1^{net}} x_{i1}^{net} + \alpha_{x_2^{net}} x_{i2}^{net}$	
FAM (incorrect outcome) - SC	$\beta_Z^{[j]} Z_i + \beta_G^{[j]} G_i + \beta_{ZG}^{[j]} Z_i G_i + \beta_{\hat{\lambda}}^{[j]} \hat{\lambda}_i^{[j]}$	$\alpha_{x_1^{ind}} x_{i1}^{ind} + \alpha_{x_2^{ind}} x_{i2}^{ind} + \alpha_{x_1^{net}} x_{i1}^{net} + \alpha_{x_2^{net}} x_{i2}^{net}$	
FAM (incorrect PS) - SC	$\beta_Z^{[j]} Z_i + \beta_G^{[j]} G_i + \beta_{ZG}^{[j]} Z_i G_i +$ $\beta_{x_1^{net}}^{[j]} x_{i1}^{net} + \beta_{x_2^{net}}^{[j]} x_{i2}^{net} + \beta_{\hat{\lambda}}^{[j]} \hat{\lambda}_i^{[j]}$	$\alpha_{x_1^{ind}} x_{i1}^{ind} + \alpha_{x_2^{ind}} x_{i2}^{ind} + \alpha_{x_1^{net}} x_{i1}^{net}$	

3.3.2 Simulation results

This section presents the results from the simulations carried out and determines some limitations of the estimation procedure laid out in Section 3.2.2. For each simulation, I generate a specific number of schools as described in the data generating section. I compare the results when considering 100 schools and 5 schools to simulate both the scenario presented by the authors, as well as a scenario more in line with the data available from the STASH intervention. Using Equation 3.4 and Equation 3.5 I calculate the true direct and spillover effects. In each simulation, the different estimates to the truth are compared to determine the bias in the estimation procedure.

Bootstrap estimates

To get uncertainty bounds we use the egocentric bootstrap method. As described in the appendix of Forastiere et al. (2020), the egocentric bootstrap method consists of drawing independent samples with replacement from the original sample of individuals, with the same number of observations. Every "observation" in a re-sample carries with it its individual-level covariates, as well as its neighbourhood-related variables (the exposure from other units being treated, as well as the neighbourhood-related covariates), even when the units in its neighbourhoods that contributed to these variables were not included in the re-sample. Following Kolaczyk (2009), egocentric sampling is valid because the observed data were obtained following the same procedure, namely, there is a chance we do not observe the entire school network but we sample the individuals at every school, ask them about their connections in the school, and then match with the names of other individuals. For each simulation we drew 500 re-samples using this bootstrap method.

Bias and variation

In the simulation study, my primary interest is the behaviour of the bias from the estimation, defined as the difference between the calculated estimates and the truth. Additionally, we want to know the level of uncertainty around that calculation. For this we compute the average standard deviation from the bootstrapped estimates. Table 3.2 shows the difference between carrying out the simulations with 100 schools and with 5 schools for the original data generating process when estimating the direct effect in three 2-column panels: one for the bias, and two to describe the level of uncertainty from the estimates. The first panel shows the average bias (calculated as the truth minus the estimate) from estimating the direct effect. In the second panel I use the bootstrap to calculate the standard deviation of the direct effect for each simulation, and then average across simulations. The final panel shows the standard deviation of the bias estimates, also referred to as the population standard deviation. One way of looking at the data generating process is as if it were the population we want to draw inference on. This means that each simulation is a draw from the population, and the information in column 3 is the truth we use to determine how closely the average bootstrapped standard errors (column 2) are to the empirical variability. Table 3.5 shows the same information for the simulations based on the data generating process that stems from Equation 3.3.

Following Forastiere et al. (2020), we use an egocentric bootstrap method to produce confidence intervals around every estimate. The observations are resampled 500 times to create a confidence intervals. The confidence intervals are calculated using the a Wald type of confidence interval. This approach from Wasserman (2004), considered the *normal* interval, produces estimates according to the following formula:

$$T_n \pm z_{\alpha/2}\hat{\text{se}}_{\text{boot}}$$

where T_n is the estimated statistic from the original sample, and $\hat{\text{se}}_{\text{boot}}$ is the bootstrap standard error. $z_{\frac{\alpha}{2}}$ represents the level of confidence we are interested in capturing, i.e. 95%, which makes $z_{\frac{\alpha}{2}} = 1.96$.

Following the logic of Table 3.2, Tables 3.3 and 3.4 show the same column configuration for the spillover effect for untreated ($Z = 0$) and treated ($Z = 1$) cases, respectively for the original data generating method.

From Tables 3.2, 3.3, and 3.4 we can see that the bias from the estimation is similar with different sample sizes, but the uncertainty around that estimate increases with smaller sample size. The FAM methodology from Forastiere et al. produces estimates with little bias when we know the propensity score model for the individual and neighbourhood treatments as well as the true outcome model. However, in the case of underspecification of the outcome model (i.e. FAM (incorrect outcome)), there is evidence of considerable bias in estimation of the spillover effects. I

Table 3.2: Bias and standard error (SE) in estimation of direct effect. Comparison: 100 vs 5 schools across 100 simulation replicates. True value of direct effect: 100 schools = 10.5; 5 schools = 10.45

Estimation method	Bias		Standard Error			
			Average Bootstrap SE		Monte Carlo SE	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	-3.93	-2.95	0.23	1.07	0.21	1.37
Adjusted	-3.88	-3.38	0.17	0.72	0.16	0.79
Splines (correct)	-0.06	-0.32	0.22	0.94	0.18	0.85
Splines (incorrect PS)	1.87	1.52	0.24	1.07	0.16	1.21
Splines (correct) - Subclasses	0.07	-0.30	0.21	2.27	0.15	0.75
Splines (incorrect PS) - Subclasses	2.32	1.78	0.24	1.53	0.19	1.30
FAM (correct)	-0.03	0.03	0.17	0.71	0.06	0.48
FAM (incorrect outcome)	1.25	0.83	0.27	1.24	0.19	0.70
FAM (incorrect PS)	-1.61	-1.94	0.14	0.75	0.16	0.67

Table 3.3: Bias and standard error (SE) in estimation of spillover effect when ($Z = 0$). Comparison 100 vs 5 schools across 100 simulation replicates. True value of direct effect: 100 schools = 3.55; 5 schools = 3.28

Estimation method	Bias		Standard Error			
			Average Bootstrap SE		Monte Carlo SE	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	1.77	1.38	0.10	0.42	0.13	0.50
Adjusted	-1.19	-1.03	0.05	0.23	0.05	0.23
Splines (correct)	-0.04	-0.37	0.12	0.45	0.12	0.48
Splines (incorrect PS)	-0.04	-0.34	0.11	0.44	0.12	0.45
Splines (correct) - Subclasses	0.14	-0.21	0.13	0.64	0.14	0.48
Splines (incorrect PS) - Subclasses	-0.07	-0.17	0.13	1.74	0.12	0.50
FAM (correct)	-0.10	-0.00	0.06	0.35	0.08	0.33
FAM (incorrect outcome)	3.20	2.92	0.13	0.65	0.18	0.52
FAM (incorrect PS)	-0.18	0.02	0.07	0.34	0.07	0.39

highlight this result because it suggests that the FAM approach does not operate in the same way as traditional propensity score regression. This is, it is not enough to assume the propensity score is modelled correctly and to include any treatment-covariate interactions properly; omitting key variables in the outcome model has a considerable impact on the bias. Removing x_1^{net} and x_2^{net} does not result in unmeasured confounding because we are including it the individual and neighbourhood propensity scores, and both of these are used in the outcome model through the subclasses and $\hat{\lambda}$, respectively. In the case of misspecification of the individual treatment model (FAM (incorrect

Table 3.4: Bias and standard error (SE) in estimation of spillover effect when ($Z = 1$) Comparison: 100 vs 5 schools across 100 simulation replicates. True value of direct effect: 100 schools = 4.08; 5 schools = 3.74

Estimation method	Bias		Standard Error			
			Average Bootstrap SE		Monte Carlo SE	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	1.93	1.41	0.14	0.58	0.18	0.88
Adjusted	-1.09	-0.94	0.08	0.36	0.10	0.33
Splines (correct)	0.81	0.38	0.17	0.65	0.13	0.85
Splines (incorrect PS)	0.32	-0.00	0.16	0.59	0.14	0.75
Splines (correct) - Subclasses	0.33	-0.03	0.21	0.82	0.15	0.62
Splines (incorrect PS) - Subclasses	0.04	-0.16	0.17	0.73	0.12	0.76
FAM (correct)	-0.09	-0.13	0.09	0.43	0.10	0.43
FAM (incorrect outcome)	3.46	2.74	0.22	1.01	0.22	0.87
FAM (incorrect PS)	-0.25	-0.16	0.08	0.41	0.09	0.49

PS)), there appears to be more bias when calculating the direct effect when compared to the FAM (correct) model.

Using splines regressions (with and without subclasses) produce similar estimates as the FAM method when the right individual propensity score model is assumed for the direct effect, as well as for the spillover effect when $Z = 0$. When we consider the spillover effect when $Z = 1$ (Table 3.4), the bias is reduced when we calculate the potential outcomes within subclass. However, in all cases, we see that using our proposed flexible regression model produces less biased results than the comparable FAM model (FAM (incorrect outcome)). The bias of the estimates decreases when we use subclasses compared to when do not use subclasses, both with the correct and the incorrect individual propensity score models. These results are important because they show that we can recover the relatively unbiased estimates when we have the correct individual propensity score model by using the cubic splines methodology instead of the FAM methodology, and that these results can be improved by using the subclassification methodology. However, we cannot recover the true estimates when the individual propensity score model is misspecified, regardless of how we estimate the outcome model.

In all tables we can see that the uncertainty around the estimates increases when the sample size is reduced from 100 schools to five schools only. This is particularly noticeable when comparing the standard errors (second panel) of the ‘Splines (correct)’ with those of the ‘Splines (correct) - Subclasses’, and of the ‘Splines (correct)’ with that of the ‘Splines (incorrect PS) - Subclasses’.

The Monte Carlo standard errors are the same or smaller when compared to the average bootstrap standard error across simulations. The results considering no-spillover effects (i.e., $\delta = 0$) in 3.2 in tables A.2, A.3, and A.4 are produced to check the robustness of our results in a different scenario and presented in the Appendix.

Alternative data-generation with smooth outcome function

As mentioned in Section 3.3, I designed an alternative data generating process, mostly determined by the change in Equation 3.2 from a using the individual propensity score in a step function with break at 0.7, to a cubic function (see Equation 3.3). I present here the results as discussed in the previous section, for the simulations that used this function to generate the outcome. Table 3.5, 3.6 and 3.7 follow the same format as their counterparts from the previous section.

Comparing Table 3.5 with Table 3.2 suggests that my proposed flexible regression outperforms the FAM approach assuming the individual propensity score is correct. These results remain valid for the estimation of the spillover effect when $Z = 0$. In the case of the spillover when $Z = 1$, the FAM methodology outperforms the flexible regression, assuming that it is using the correct individual propensity score model. The takeaway from these results is similar to the one from the previous section: when we have the correct individual propensity score model, flexible regression seems to outperform the FAM methodology in most estimation routines without the need have perfect information about the outcome model. Deviations from the individual propensity score affect both methodologies in a similar way.

3.3.3 Coverage

Coverage is the average of the number of times the true estimate lies in the confidence interval given by +/- 1.96 the bootstrap SE for every simulation. From Table 3.8 we can see that the estimates recovered using our flexible regression approach, both with and without subclasses, and the "FAM (correct)" estimates, perform similarly for the direct effect in terms of coverage. When it comes to coverage of the spillover effects, the increased performance of the method with subclasses, shows that the it is the subclasses, and not necessarily the used outcome model, the part of the estimation routine that improves the results. Considering the results from Table 3.2, 3.3 and

Table 3.5: Bias and standard error (SE) in estimation of direct effect. Comparison 100 vs 5 schools across 100 simulation replicates. Alternative data generating model. True value of direct effect: 100 schools = 10.5; 5 schools = 10.45

Estimation method	Bias		Standard Error			
			Average Bootstrap SE		Monte Carlo SE	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	-3.03	-2.00	0.22	1.03	0.19	1.17
Adjusted	-3.01	-2.46	0.13	0.56	0.13	0.76
Splines (correct)	-0.02	-0.06	0.20	0.87	0.22	0.68
Splines (incorrect PS)	2.20	1.98	0.22	1.00	0.13	0.84
Splines (correct) - Subclasses	0.08	-0.16	0.20	2.20	0.18	0.65
Splines (incorrect PS) - Subclasses	2.48	2.23	0.22	1.85	0.16	1.00
FAM (correct)	-1.09	-0.86	0.08	0.47	0.09	0.73
FAM (incorrect outcome)	0.21	0.27	0.21	1.11	0.16	0.62
FAM (incorrect PS)	-1.56	-1.61	0.09	0.48	0.11	0.44

Table 3.6: Bias and standard error (SE) in estimation of spillover effect when ($Z = 0$). Comparison 100 vs 5 schools across 100 simulation replicates. Alternative data generating model. True value of direct effect: 100 schools = 3.54; 5 schools = 3.27

Estimation method	Bias		Standard Error			
			Average Bootstrap SE		Monte Carlo SE	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	1.77	1.39	0.10	0.42	0.18	0.47
Adjusted	-1.16	-1.04	0.05	0.22	0.06	0.23
Splines (correct)	-0.00	-0.34	0.12	0.46	0.14	0.46
Splines (incorrect PS)	-0.01	-0.32	0.11	0.45	0.13	0.44
Splines (correct) - Subclasses	0.11	-0.17	0.13	0.64	0.15	0.47
Splines (incorrect PS) - Subclasses	-0.06	-0.18	0.12	2.03	0.14	0.43
FAM (correct)	-0.14	-0.01	0.06	0.32	0.10	0.32
FAM (incorrect outcome)	3.12	2.95	0.13	0.62	0.22	0.51
FAM (incorrect PS)	-0.18	0.01	0.07	0.32	0.09	0.31

3.4, we can say that the higher coverage in the case of the small sample size is due to the comparatively larger standard errors, and not because of higher estimator accuracy. The lower coverage is correlated with the accuracy of the estimator, which explains why the values that have less accuracy and high sample sizes have a coverage of 0 most of the time. In other words, the poor coverage occurs because the bias does not decrease with increasing sample size, but standard errors do. This also translates into the fact that Table 3.9 shows that when we use our alternative data generating

Table 3.7: Bias and standard error (SE) in estimation of spillover effect when ($Z = 1$). Comparison 100 vs 5 schools across 100 simulation replicates. Alternative data generating model. True value of direct effect: 100 schools = 4.07; 5 schools = 3.75

Estimation method	Bias		Standard Error			
			Average Bootstrap SE		Monte Carlo SE	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	1.83	1.31	0.13	0.55	0.23	0.89
Adjusted	-1.16	-1.03	0.07	0.33	0.13	0.34
Splines (correct)	0.75	0.24	0.17	0.63	0.17	0.73
Splines (incorrect PS)	0.31	-0.08	0.15	0.58	0.16	0.66
Splines (correct) - Subclasses	0.37	0.05	0.20	0.82	0.18	0.69
Splines (incorrect PS) - Subclasses	0.15	-0.19	0.17	0.71	0.18	0.74
FAM (correct)	-0.04	-0.08	0.08	0.37	0.14	0.43
FAM (incorrect outcome)	3.47	2.86	0.21	1.01	0.23	0.89
FAM (incorrect PS)	-0.16	-0.15	0.08	0.37	0.13	0.47

process, the splines estimators outperform the FAM estimators in terms of coverage for all three quantities of interest.

Table 3.8: Nominal 95% coverage of Wald-type confidence intervals with standard errors based on 500 bootstrap resamples for each of the 100 simulations for the original data generating method

Estimation method	Main effect		Spillover effect ($Z = 0$)		Spillover effect ($Z = 1$)	
	Simulation setting		Simulation setting		Simulation setting	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	0.00	0.26	0.00	0.00	0.00	0.28
Adjusted	0.00	0.00	0.00	0.00	0.00	0.08
Splines (correct)	0.98	1.00	0.98	0.80	0.00	0.98
Splines (incorrect PS)	0.00	0.74	0.96	0.80	0.54	1.00
Splines (correct) - Subclasses	0.96	1.00	0.78	1.00	0.82	1.00
Splines (incorrect PS) - Subclasses	0.00	0.80	0.98	1.00	1.00	1.00
FAM (correct)	1.00	1.00	0.94	1.00	1.00	1.00
FAM (incorrect outcome)	0.00	1.00	0.00	0.02	0.00	0.22
FAM (incorrect PS)	0.00	0.14	0.02	1.00	0.08	1.00

3.4 STASH

The Sexually Transmitted infections And Sexual Health feasibility study was carried out by the Social and Public Health Sciences Unit of the University of Glasgow in six schools throughout

Table 3.9: Nominal 95% coverage of Wald-type confidence intervals with standard errors based on 500 bootstrap resamples for each of the 100 simulations for the alternative Data Generation Process

Estimation method	Main effect Simulation setting		Spillover effect ($Z = 0$) Simulation setting		Spillover effect ($Z = 1$) Simulation setting	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	0.00	0.54	0.00	0.00	0.00	0.30
Adjusted	0.00	0.00	0.00	0.00	0.00	0.02
Splines (correct)	0.84	0.98	0.98	0.78	0.00	1.00
Splines (incorrect PS)	0.00	0.70	1.00	0.76	0.56	1.00
Splines (correct) - Subclasses	0.82	1.00	0.96	1.00	0.60	1.00
Splines (incorrect PS) - Subclasses	0.00	0.54	0.98	1.00	0.92	0.98
FAM (correct)	0.00	0.34	0.28	1.00	1.00	1.00
FAM (incorrect outcome)	0.88	1.00	0.00	0.00	0.00	0.12
FAM (incorrect PS)	0.00	0.06	0.00	1.00	0.52	1.00

Scotland for students between the ages of 14 and 16. The intervention was designed to encourage influential students, chosen by their peers, to start conversations with other students about sexual health on social media and face to face.

Figure 3-1 shows all six schools, its students, the connections between the students, and how some students are peer supporters (blue), exposed to treatment (yellow), peer supporters who were also exposed to treatment (green), and pure controls (white). Table A.6 (in the Appendix) shows a summary of relevant variables from the STASH dataset. In this analysis, and following Forsyth et al. (2018), we consider as outcome of the intervention an index created by adding up the level of confidence students have in answering three questions related to sexual behaviours: “how confident are you to get condoms on your own?”, “how confident are you to put a condom on yourself or a partner?”, and “how confident are you to refuse to have sexual intercourse if they won’t use a condom?”. I refer to this outcome variable as the Confidence in Sexual Health (CSH) index throughout this section; it ranges from 3 to 15 with a mean of 10.96 and a median of 11.

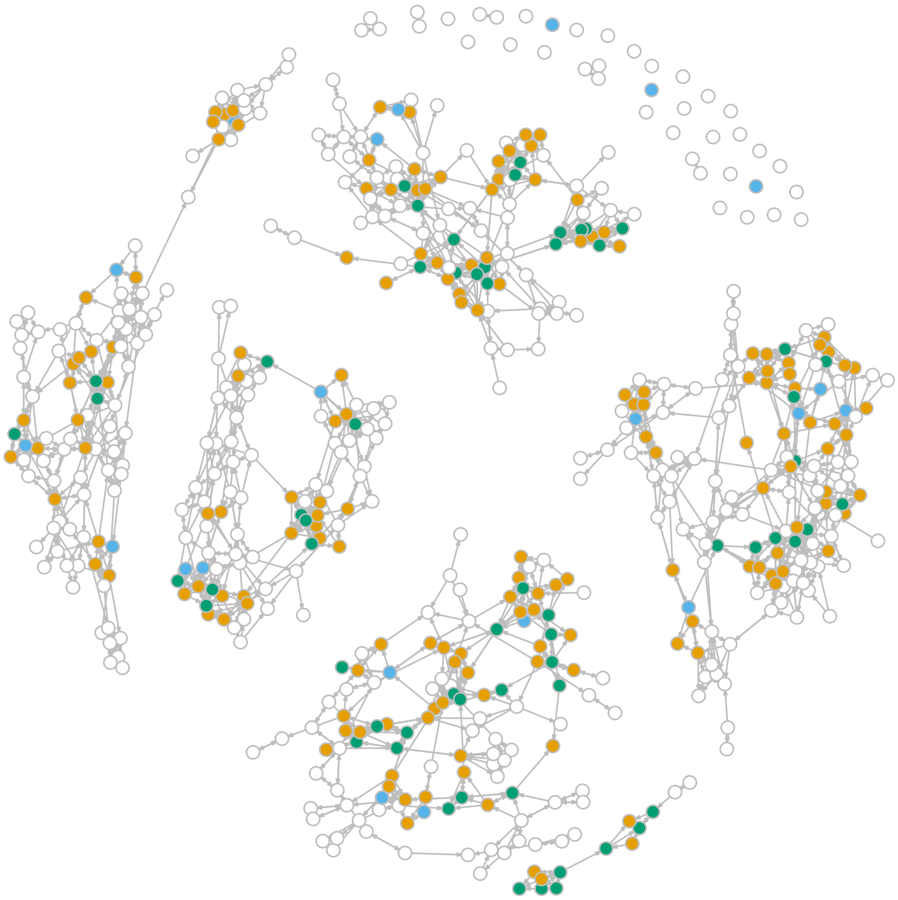


Figure 3-1: Social network of students in STASH schools indicating treatment (blue), exposition to treatment (yellow), treatment and also being exposed to treatment (green), and controls (white).

In addition to the individual covariates, we are interested in the distribution of network variables for the students in the schools. The network information gathered by the researchers is that of admiration or "looking up to", capped at 6 in the questionnaire, which has a clear effect in the distribution of the "out-degree" variable (as seen Figure A-3 in the appendix). Most students have between one and four other students who look up to them, with fewer being looked up to by 6 or more.

For this particular application the network variable of interest is the out-degree. This is in contrast to using in-degree as the network variable of interest for the simulation in Section 3.3. In both cases I am trying to capture how the intervention might flow from one individual to the next. In the particular case of STASH, out-degree is important because students look up to peer supporters, which assumes they are more inclined to follow their actions than if we considered the incoming connections. For the purpose of this section, we consider a peer supporter a treated individual, and an individual who is in the neighbourhood of that peer supporter to be an exposed individual.

3.4.1 Model

The previous section explained how Forastiere et al. (2020) estimate direct and spillover effects using their generalised propensity score methodology. This section shows my modelling of the individual and neighbourhood propensity scores for the STASH dataset. The models here are used for the FAM and for the flexible regression approaches described in Section 3.3. Let us begin with the individual propensity score.

- Individual Propensity Score. Following an initial survey which establishes an admiration network between the students at every treated school, approximately the top 25% of the in-degree distribution (those who were considered most looked up to by their peers) were asked if they wanted to participate in the intervention as peer-supporters. Half of these were randomly selected to be part of the treatment group and trained as peer supporters. I refer to the fact of being nominated as a peer supporter as being treated (using Z as the variable that indicates treatment). The questions that determined admiration were designed to encourage nominations to peers that were close and important for every student.

  The function that determines the propensity of being assigned to treatment, $\phi(z; x^z)$ depends on the observed values for the set of variables X^z. I assume that the variables that determine the propensity of being assigned to treatment are: in-degree and whether the individual has had sex before or not.

- Neighbourhood Propensity Score. The nominated peer supporters were tasked with getting in touch with their connections via Facebook and off-line, discussing the main STASH talking points and also advertising local sources of information on the topic of sexual and reproductive health. I refer to this interference in treatment as the neighbourhood effect, G.

The function that determines the propensity of being exposed to treatment by a treated peer, $\lambda(g; z, x^g)$ depends on the observed values for the set of variables X^g. I assume that the variables are the treatment, out-degree, and level of self-esteem.

- The outcome variable we are using to measure the effect of the intervention is the CSH index defined in the beginning of this section. The battery of questions used to create this index are considered the main outcome variable by the researchers who carried out the intervention. Considering the delicate nature of the topic and how relatively uncomfortable teenagers might feel with it, an index, and not a specific variable or variables is best suited to determine the effectiveness of the intervention.

Estimation I estimate the results using a subsection of the methodologies described in Section 3.3. Since we believe the specification of the individual treatment model, as well as the neighbourhood treatment model to be right, we do not need to estimate the direct and spillover effects using different specifications. In terms of the FAM methodology, we suggest an outcome model to be used in step 4.b from Section 3.2, with the following functional form:

$$CSH \sim \beta_{gender}\text{gender} + \beta_z z + \beta_g g + \beta_{\hat{\phi}}\hat{\phi} + \beta_{\hat{\lambda}}\hat{\lambda} + \beta_{\hat{\lambda},z}\hat{\lambda}z + \beta_{\hat{\lambda}^2}\hat{\lambda}^2. \tag{3.7}$$

This is, the outcome depends on the treatment, how much exposure they received to the treatment via their peers, the propensity of to be assigned to treatment, the propensity to be exposed to treatment, the interaction between the propensity to be exposed to treatment and treatment itself, and the square of the propensity of being exposed to treatment (as suggested by Hirano and Imbens, 2005 and Forastiere et al., 2020, to include some higher order terms that might affect the dose in the dose-response function). I include the gender (assigned at birth) of the individual as a control for the outcome model.

Regarding the flexible regression approach, we follow the procedure described in Section 3.3, using the same individual and neighbourhood propensity scores as the FAM methodology, and the following outcome model:

$$W_i(Z_i, G_i) \sim \beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i G_i + \beta_{\hat{\phi}} \text{spline}(\hat{\phi}_i) + \beta_{\hat{\lambda}} \text{spline}(\hat{\lambda}_i)$$

in the case for no subclassification, and

$$W_i^{[j]}(Z_i, G_i) \sim \beta_Z Z_i + \beta_G G_i + \beta_{ZG} Z_i G_i + \beta_{\hat{\phi}} \text{spline}(\hat{\phi}_i) + \beta_{\hat{\lambda}} \text{spline}(\hat{\lambda}_i^{[j]})$$

in the case where the data are split into subclasses.

3.4.2 Results

The results presented in the panel of Table 3.10 indicate a higher level in the aggregate measure of confidence in sexual health related issues for the intervened individuals that is not greatly affected by confounding. As explained in Section 3.4, the treated individuals are the peers selected as peer supporters. The FAM methodology, which is supposed to filter out the effect of the spillover from the causal direct effect of the intervention, estimates the latter to be around 1 CSH index point. The second and third panels of Table 3.10 show that the spillover had little to no effect on the exposed or unexposed individuals, even though there was perhaps some slight suggestion in the data that the exposed adolescents were slightly more susceptible to messaging from their treated peers than those teens chosen to be controls.

In this relatively small, feasibility study, any reduction in bias from the estimation of direct and spillover effects may be offset by increased variability, as we saw with the large standard errors in Tables 3.2 to 3.7. The approach nevertheless serves to provide confidence in the findings, especially in light of the recent publication by Hirvonen et al. (2021) suggesting that the level of influence intended to change the behaviour of exposed students in the school was not achieved. The authors mention a failure in the implementation of the interference mechanism. Specifically, the way the peer supporters were expected to share or propagate the intervention over to their peers was by getting in touch with them on online social forums. Their qualitative study of the trial indicates that both peer supporters and their intended audience did not interact as actively and often as desired. Hirvonen et al. (2021) suggest that this was one of the primary reasons the treatment did not show differences in the composite CSH index for the individuals that were not selected as peer supporters.

Similarly, as discussed in Section 3.3, it is important to consider the uncertainty of the estimated values. I use the same methodology described in Section 3.3 to generate the standard errors around the point estimates.

Table 3.10: STASH - main and spillover effects of the intervention. Total number of students: 605; 6 schools.

Estimation method	Main effect		Spillover Effect			
			$Z = 0$		$Z = 1$	
	Estimate	Bootstrap CI	Estimate	Bootstrap CI	Estimate	Bootstrap CI
Unadjusted	1.04	[0.37, 1.74]	0.03	[-0.14, 0.21]	0.13	[-0.15, 0.40]
Adjusted	0.99	[0.32, 1.69]	0.01	[-0.15, 0.19]	0.09	[-0.20, 0.36]
Splines	1.09	[0.30, 1.88]	-0.01	[-0.40, 0.38]	-0.06	[-0.60, 0.37]
FAM	0.95	[0.16, 1.73]	0.02	[-0.34, 0.21]	0.22	[-0.49, 0.61]

3.5 Discussion

I show, using an extension of the methodological tools proposed by Forastiere et al. (2020), that the Sexually Transmitted infections And Sexual Health feasibility study carried out by the University of Glasgow's Social and Public Health Sciences Unit did not seem to have the intended effect of disseminating its treatment through social connections onto the individuals that were not intervened. I consider the Confidence in Sexual Health index as our main outcome variable, according to the study's report (Forsyth et al., 2018), and find that the intervention increased the value of the CSH index for treated individuals as it was expected, but not for individuals that were indirectly exposed to treatment, as it was desired.

My results are in line with a recent publication (Hirvonen et al., 2021) by part of the team that performed the feasibility study stating that the message sharing carried out by peer supporters was hindered by the irregular engagement with the platform the students were using. In other words, the "spillover" part of the intervention did not work as intended, essentially eliminating the transmission of the treatment to peers originally considered to be controls.

The simulation analysis shows that the methodology proposed by Forastiere et al. (2020) requires knowledge of the true model for individual treatment as well as the true outcome model in order to produce the desired results. I further show that underspecifying the outcome model produces bias in the estimation of the spillover effects. In contrast, using the flexible regression methodology with the individual and neighbourhood treatment estimated propensity score in the outcome model, bias in the estimation of the spillover effect is reduced. This result is maintained when we consider a different data generating process that does not neatly fit with subclass (i.e. a step function), but rather has a more general shape. The flexible regression generates results with different degrees of bias depending on whether we use the methodology with subclasses or not, as explained in Subsection 3.2.3. Relevant to our empirical case study, I show that the uncertainty around the estimates increases when the sample size is small.

Note we do not know what the exact true neighbourhood model is, nor we study the performance of the estimators in relation to different specifications of the neighbourhood model. This is because spillover is determined by the number of connections each unit has, and this is in turn determined by the structure of the network. Future research should focus on looking at different network generating processes, i.e. how the connections between the individuals are formed. For example, networks that exhibit more homophily (connections are formed based on existing attributes) can generate different influence patterns than networks where connections are, for example, produced at random. This is particularly important for the case in which the variable that drives the homophily pattern also drives the selection of individuals into behaviour. For the STASH analysis, it is possible that gender homophily, or some other kind of homophily producing groupings of similar individuals, affects the prediction of neighbourhood effects biasing the results. However, without more experimentation or a different simulation analysis than the presented here, there is no way of knowing how homophily can affect the estimation.

Additionally, the limit of the maximum number of connections to 5 will likely impose a limit to how much interference is captured by the model. In relation to the analysis of the STASH intervention, it could mean an underestimation of the actual amount of interference, since students have more connections that were disclosed, albeit less strong ones. Note that this is not a problem in the simulation analysis because the interference is introduced with the restriction on the number of connections already in place. More research is needed to determine the how limiting the number of connections, especially when these are weaker connections, affects the estimation of spillover effects.

One of the limitations of my methodological approach that uses splines to estimate the direct and spillover effects, is that the uncertainty around the point estimate is larger when compared to the FAM methodology proposed by Forastiere et al. (2020). Additionally, as shown in the results tables in 3.3, the uncertainty with the egocentric bootstrap approach around the point estimates increases with a small sample size like that of the STASH study. This is particularly noticeable with smaller sample sizes. In general this highlights the need for more research in understanding appropriate variance estimators in the context of interdependent data.

Chapter 4

Identification of social influence in bipartite cascades of political behaviour

4.1 Bipartite event cascades and political behaviour

Political actors make interdependent choices. Their behavior is often embedded in a system of behavior displayed by other actors. In international relations, for example, states declare war on countries or forge alliances based on the war and alliance portfolios of other countries (Cranmer et al., 2012). They ratify international treaties in relation to the ratification behavior of other states (Campbell et al., 2019). In domestic politics, legislators may vote on the adoption of laws or cosponsor bills conditional on past voting or sponsorship behavior of other legislators (Ringe et al., 2013; Fowler, 2006). Interest groups and government agencies signal their policy beliefs to decision makers and the public through public debate. They react to the adoption of policy beliefs by other actors in the pursuit of influencing the outcome of the policy process (Leifeld, 2017). In these examples of iterative and interactive politics, actors from a well-defined population – legislators in a legislature, countries in the international system, and interest groups in a policy process – display publicly observable ties to categorical choices in a temporal sequence – treaties, international organizations, bills, and policy beliefs.

Political science tries to explain events in these sequences: Why do countries attack a certain target? Why do legislators cosponsor a certain bill? The behavior of each actor at any time point depends, in part, on the observable behavior by others earlier in the sequence. Hence the data can be modeled as temporal bipartite networks, in which a first-mode node (the actor or "sender") develops a tie to a second-mode node (the categorical behavior or "receiver") conditional on the past sequence of events. Such kinds of networks are also known as bipartite behavior cascades (Kleinberg, 2007). Until recently, political science focused on explanations of ties in bipartite behavior cascades using single nodes on one mode, or node pairs in a two-mode networks. Ties were assumed to be exogenously determined and modeled as a function of characteristics of the sender, the receiver, the sender–receiver pair, and/or time, often in a time-series cross-sectional modeling approach (Franzese and Hays, 2007, 144).

Different parts of the discipline have now developed a more nuanced understanding of endogenous processes playing a role in bipartite behavior cascades. One active field of research is Political

Networks (Victor et al., 2017). It posits that dependencies between observations matter theoretically and statistically, and their omission would lead to omitted variable bias (Cranmer et al., 2017b). Temporal models include extensions of exponential random graph models or latent space models to panel data, and relational event models for modeling dependence in event history data. Another area recognizing the interdependence of behavior cascades is Policy Diffusion, which disentangles different endogenous mechanisms by which behavior spreads across actors over the course of a cascade (Shipan and Volden, 2008; Desmarais et al., 2015). A third research area that recognizes interdependent behavior of actors over time is the literature on Spatio-Temporal Autoregressive Models. Here, the behavior of an actor depends on the contemporaneous and past behavior of other actors in a panel of cross-sectional observations, and the workhorse model is the spatial autoregressive model (SAR) and its extensions (Franzese and Hays, 2007).

In all three areas of inquiry, the goal of causal identification of endogenous mechanisms has grown in importance in recent years. The biggest challenge stems from the observation that the temporally unfolding dependence of behavior between actors is often compatible with multiple plausible causal pathways. Separating these pathways has proven difficult. One prominent causal identification problem is the confounding between social influence and homophily among actors (Aral et al., 2009). Social influence means that one actor influences another actor's behavior in a cascade. Homophily usually means that actors develop network ties because they share predispositions. Here I take it to mean that actors display similar cascade behavior because they were similar to begin with, and there is no endogeneity in the choices.

Both causal mechanisms frequently co-occur in political behavior. For example, it is hard to separate the effect of states learning from politically similar states in a policy diffusion study from the effect of shared political characteristics independently leading to similar policy adoption behavior. How can we disentangle them in event-based cascades of behavior in order to explain politics more effectively? That is, how can we distinguish social influence statistically from the effect of prior similarity of units in bipartite network cascades? In the present chapter, I contribute to exploring this problem by proposing a shuffle test and evaluating its effectiveness and boundary conditions.

4.2 Confounding in cascades of political behavior

The problem of confounding in bipartite behavior cascades is an aggravated version of the problem of causal confounding in temporal networks more generally. In this more general setting, one can observe both the network and the behavior of nodes ("attributes") over time. The task is to disentangle the causal relationships between network formation and the behavior of the nodes in the network. Do network ties cause individual behavior (e. g., through diffusion, contagion, or imitation), or does behavior cause tie formation or dissolution (e. g., through homophily, i. e., similarities in behavior leading to network ties) (Aral et al., 2009)? This question is highly consequential for understanding many real-world outcomes, such as the complex relationships between war, peace, conflicts, alliances, democracy, and autocracy in the study of international relations (Gleditsch and Ward, 2006; Lee Ray, 2013). Yet, statistical identification of the causal direction is one of the hardest problems in network science because "homophily and contagion are generically confounded in observational network studies" (Shalizi and Thomas, 2011).

Adding to the severity of the challenge, in many real-world applications the underlying social network is unobserved. In such cases, the questions of whether the behavior of one node in a network influences the behavior of another node in the network over time and what covariates predict this process must be answered solely based on the observed sequence of behavior of the units, without knowledge of the actual underlying social network (Gomez-Rodriguez et al., 2012, 2013; Rodriguez et al., 2014; Desmarais et al., 2015; Campbell et al., 2019; Marrs et al., 2019). I refer to data organised in a two-mode as call bipartite event cascades. Examples abound: Gomez-Rodriguez et al. (2012) inferred the latent diffusion paths among blogs (Mode I) through the sequence of shared memes (Mode II); Desmarais et al. (2015) inferred the latent diffusion pathways among U.S. states (Mode I) through the sequence of policy adoptions (Mode II); Campbell et al. (2019) modeled the ratification sequence of international environmental treaties (Mode II) by countries (Mode I) as a function of similar levels of economic development (homophily covariate); and Malang et al. (2019) explained the sequence of expressions of negative opinions by national parliaments in the European Union (Mode I) on legislative bills (Mode II) as a function of ideological similarities among parliaments (homophily covariate).

However, sequences of behavior can either be governed by nodes adopting other nodes' behavior over time in the sequence ("contagion" or "diffusion"), or some prior similarity variable can structure the observed behavior, without any influence over time taking place. As Anagnostopoulos et al. (2008, 2) put it, under this similarity regime, "the probability that an individual is active can be affected by whether their friends become active, but not by when they become active." For example, do states in the international system independently ratify the same treaties over time because they have similar geopolitical resources and interests, allies etc. (i. e., joint prior variables), or do they ratify the same treaties over time because one country's adoption is triggered by the recent adoption of the same treaty by another country that had similar geopolitical resources and interests or was an ally (i. e., contagion or diffusion)? The literature on bipartite event sequences tends to focus on inferring (Gomez-Rodriguez et al., 2012, 2013; Rodriguez et al., 2014; Malang and Leifeld, 2021; Desmarais et al., 2015) or explaining (Campbell et al., 2019; Marrs et al., 2019; Malang et al., 2019) latent pathways of influence and to ignore the possibility that prior similarities may lead to similar rates of behavior. For example, the fact that two nodes share a similar ideology may lead both of them to adopt a behavior rather quickly (or often) – but independently of each other. How can one ensure that any dyadic explanations of shared behavior capture influence/contagion/diffusion and not antecedent similarities leading to similar rates of independent behavior?

A powerful modeling framework for bipartite event sequences like these is the relational event model (REM) (Butts, 2008). The REM is a network model that explains the time until a network tie occurs as a function of covariates and endogenous properties of the past network sequence, such as prominence (emergence of key nodes in the network), inertia (the tendency to keep connecting to a previously connected node), and other sufficient statistics for capturing network endogeneity. In a bipartite setting, the REM can capture the role of dyadic covariates between actors for triggering ties to second-mode behavioral nodes as a homophily effect. A more similar dyadic value between two actors on the first-mode dyadic covariate (e. g., two actors have the same ideology or party affiliation), coupled with recent relevant behavior ties of the peer actor (e. g., the other actor engaged in behavior *B* recently), increases one's hazard of behavior ties (e. g., to *B*), where hazard is to be understood in the survival analysis sense. The bipartite REM is an attractive model for many applications in political science. However, just like related statistical approaches, the bipartite REM cannot per se distinguish between prior similarity and contagion in these homophily effects.

To explore this further, in this chapter I present the following contributions: (1) A simulation model is presented that can create bipartite event sequences where an underlying dyadic variable causes new bipartite events through either influence or prior similarity or a linear combination of both. (2) Given these simulations, I illustrate how similarity and contagion can be temporally confounded in empirical applications of relational event models to bipartite event sequences with homophily effects. (3) A shuffle test, which can distinguish between the two mechanisms in empirical applications. (4) Using simulations, I delineate the scope conditions under which the shuffle test permits identification and describe which conditions would render the test meaningless for distinguishing between similarity and influence in empirical applications. And (5) I illustrate the efficacy of the approach using a toy example on international environmental treaty ratification.

4.3 Influence versus similarity in bipartite sequences

The homophily statistic in Equation 2.21 captures the influence of recently observed congruent behavior by compatible other senders. It is tempting to interpret a significant coefficient as proof of social influence over the event sequence. However, in some cases it may not be influence over the course of the event sequence, where senders "learn" from other senders, but rather the prior, time-invariant attribute similarity that leads independently to similar behavior.

To illustrate this point, this section presents a simulation model that can create artificial bipartite event sequences i) based on social influence given a sender covariate, ii) based on time-invariant sender similarity given the covariate, or iii) any linear combination of the two. Consider the example presented in Figure 4-1. In this diagram, we show senders connecting to receivers at different points in time. The diagram shows what influence could look like in a real event sequence: Sender S_1 connects with a couple of receivers, R_1 and R_3. After sender S_2 connects with receiver R_3, S_1 follows it and connects with R_3 as well.

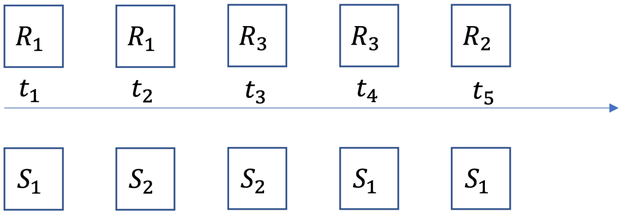


Figure 4-1: Diagram of an event sequence showing some possible evidence of influence, where sender S_1 connects with receiver R_3 after sender S_2 connected with the same receiver (R_1) after S_1 did.

In this model, I assume that there is a set of sender nodes $S = \{s_1, \ldots, s_m\}$ on the first mode and a set of receiver nodes $R = \{r_1, \ldots, r_n\}$ on the second mode. Different simulation scenarios with number of senders $m \in \{5, 10, 20, 50\}$ and number of receivers $n \in \{1, 2, 5, 10, 20\}$ are explored below. Each sender and each receiver carry an attribute, stored in vectors $\mathbf{a}_s \sim U(0, 1)$ and $\mathbf{a}_r \sim U(0, 1)$, where $q(s)$ or $q(r)$ is a function that retrieves the attribute value of any sender s or receiver r from $\mathbf{a}_s$ or $\mathbf{a}_r$. $q(s)$ and $q(r)$ are aggregation functions that compress the attributes of a sender or a receiver into a scalar which can then be used to compare with other senders or receivers.

For the receivers, the attribute represents an intrinsic quality, and for the senders, the attribute value is interpreted relative to those qualities. For example, if senders are countries and receivers are behaviors like ratifying treaty A (e. g., $q(r_1) = 0.2$) or treaty B (e. g., $q(r_2) = 0.7$), an individual value of $q(s) = 0.4$ can be interpreted as having a closer ideal point to treaty A than treaty B. (I assume here for simplicity that the preference space is one-dimensional.)

At each time point $t \in \{1, \ldots, T\}$ in an event sequence over T discrete time points, draws from the binomial distribution with parameter (n, p) determine, for each single time point, how many events take place. I choose simulation scenarios with $T \in \{50, 100, 200, 300, 500, 1000\}$ and $(n, p) = (2, 0.7)$ to allow for moderate amounts of simultaneity and a moderate probability of generating time points without any events, as is often observed empirically (e. g., Malang et al., 2019; Brandenberger, 2019). In each of these events $e \in E_T = (s, r, t)$, sender s and receiver r are chosen with probability

$$\begin{aligned} P(s_e = s, r_e = r|t, T_{1/2}, \mathbf{a}_s, \mathbf{a}_r) &= P(s_e = s|r_e = r, t, T_{1/2}, \mathbf{a}_s, \mathbf{a}_r)P(r_e = r|t, T_{1/2}, \mathbf{a}_s, \mathbf{a}_r) \\ &= P(r_e = r|s_e = s, t, T_{1/2}, \mathbf{a}_s, \mathbf{a}_r)P(s_e = s|t, T_{1/2}, \mathbf{a}_s, \mathbf{a}_r), \end{aligned} \tag{4.1}$$

where $T_{1/2}$ represents the half-life parameter, a variable introduced for temporal weighting of past events with geometric decay (see Section 2.4). This probability, which leads to the decision of which sender forms a tie to which receiver, is a finite mixture of two probability mass functions I call similarity and influence,

$$\begin{aligned} &P_{\text{decision}}(s_e = s, r_e = r|t, T_{1/2}, \mathbf{a}_s, \mathbf{a}_r) \\ &\qquad = \pi P_{\text{similarity}}(s_e = s, r_e = r|\mathbf{a}_s, \mathbf{a}_r) + (1 - \pi)P_{\text{influence}}(s_e = s, r_e = r|t, T_{1/2}, \mathbf{a}_s), \end{aligned} \tag{4.2}$$

where $\pi \in [0, 1]$ is a parameter that determines the way in which similarity and influence are blended, ranging from complete similarity ($\pi = 1$) to complete influence ($\pi = 0$) to determine the decision outcome. Note that the probability of s choosing r is independent of time under the similarity regime, but it depends on time and the pre-chosen half-life parameter ($T_{1/2}$) under the influence regime. Additionally, that the probability of s choosing r is independent of the ideological position of the receiver under the influence regime.

The similarity component of the decision is proportional to the absolute similarity (i. e., one minus the absolute difference) between the attribute of the sender and the attribute of the receiver, with a normalizing constant that sums these values over all sender and receiver combinations:

$$P_{\text{similarity}}(s_e = s, r_e = r|\mathbf{a}_s, \mathbf{a}_r) = \frac{1 - |q(s) - q(r)|}{\sum_{i=1}^{m} \sum_{j=1}^{n} \left(1 - |q(s_i) - q(r_j)|\right)} \tag{4.3}$$

This mechanism corresponds to senders choosing the same receiver as other senders who have a similar attribute value. This is made apparent by the fact that for two different s and s',

$$\begin{aligned} &1 - |q(s) - q(r)| = 1 - |q(s') - q(r)| \\ &\qquad \Leftrightarrow P_{\text{similarity}}(s_e = s, r_e = r | \mathbf{a}_s, \mathbf{a}_r) = P_{\text{similarity}}(s_e = s', r_e = r | \mathbf{a}_s, \mathbf{a}_r). \qquad (4.4) \end{aligned}$$

The similarity function is thus a homophily mechanism without regard for time, in which senders with similar prior or time-invariant attributes tend to choose similar receivers, without any social influence taking place.

In contrast, the social influence component of the decision is proportional to the similarity between the attribute of the sender and the attributes of past senders who chose the focal receiver r, weighted geometrically by time:

$$\begin{aligned} &P_{\text{influence}}(s_e = s, r_e = r | t, T_{1/2}, \mathbf{a}_s) \\ &\quad = \frac{\frac{\sum_{e^* \in E_{t-1}} [s_{e^*} \neq s][r_{e^*} = r](1 - |q(s_{e^*}) - q(s)|) w(e^*, t_e, T_{1/2})}{\sum_{e^* \in E_{t-1}} [s_{e^*} \neq s][r_{e^*} = r]}}{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\sum_{e^* \in E_{t-1}} [s_{e^*} \neq s_i][r_{e^*} = r_j] \left(1 - |q(s_{e^*}) - q(s_i)|\right) w(e^*, t_e, T_{1/2})}{\sum_{e^* \in E_{t-1}} [s_{e^*} \neq s_i][r_{e^*} = r_j]}} \qquad (4.5) \end{aligned}$$

This probability is identical to the definition of the homophily statistic defined in Equation 2.21, with a normalizing constant that sums the same values over all possible sender and receiver combinations. Note that this probability is a function of time (i. e., it depends on the past event sequence at time t) and the pre-chosen half-life parameter $T_{1/2}$, just like in the REM estimation case outlined above. This congruence of the influence mechanism and the homophily statistic enables us to recover the coefficient for social influence using simulated data with identical $T_{1/2}$ using REM estimation for discrete data. Simulation scenarios with $T_{1/2} \in \{5, 10, 20, 50\}$ is chosen to cover a broad range of cases. Because of the time dependence of the influence mechanism, simulations must proceed chronologically with each simulation step building on the simulated prior event sequence.

Table 4.1: Comparison of REM with simulated high-influence ($\pi = 0.1$) and high-similarity ($\pi = 0.9$) sequences with $m = 5$, $n = 2$, $T = 300$, $T_{1/2} = 5$, and $(n, p) = (2, 0.7)$. *** indicates significance at the 0.05 level.

	Model 1 ($\pi = 0.1$)	Model 2 ($\pi = 0.9$)
Homophily (ξ)	273.33 (23.48)***	239.02 (23.90)***
Num. events	529	541
Num. obs.	2571	2427

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Each simulation run with pre-chosen m (number of senders), n (number of receivers), T (number of time points), $B(n, p)$ (number of events per time unit), $T_{1/2}$ (half-life of past event sequence), π (relative importance of similarity vis-a-vis influence), and randomly drawn uniform values in $\mathbf{a}_s$ and $\mathbf{a}_r$ (attribute for senders and receivers) yields one artificial event sequence. $\gamma = 100$ such sequences are simulated per scenario and subsequently supplied to a REM estimation routine to recover homophily coefficients. The REM estimation is performed on the last $T - \frac{T}{10}$ time points. Discarding the first ten percent of the simulated data allows for a proportional burn-in period during which the simulations build up a sufficient history for the influence mechanism to work effectively.

REMs cannot reliably distinguish between the effects of prior, or time-invariant, similarity of attributes and social influence. I measure this using the homophily statistic (ξ) To illustrate this point, Table 4.1 shows a comparison of two estimated REMs. I simulated two sequences, one with high influence and one with high similarity, and estimated the homophily coefficient from Equation 2.21, which corresponds directly to the definition of the influence mechanism in Equation 4.5. The coefficients are both positive and significant. Consequently, the identification of the causal mechanism is likely hindered in empirical studies, where analysts may be tempted to interpret prior similarity as influence, contagion, or diffusion. In the next section, I will present a shuffle test to rectify the problem.

4.4 Directed acyclic graphs and identifiability

In this section I present diagrams representing the data-generating process for the simulations described in the main document. I highlight the relationships between the different variables using a Directed Acyclic Graph (DAG), which indicates the flow of causal influence between variables. The DAG is directed because there is a directional flow from one variable to the next and acyclic because causality cannot simultaneously flow back to prior variables on a path. Causal identification requires not just the absence of cyclical edges in the graph, but also the absence of plausible unmodeled backdoors (confounders) (Pearl, 1995). Hence I present extensions of the model below, where the possibility of factors that may jeopardize causal identification is discussed further and empirical remedies are suggested. Table 4.2 describes the variables and constants in the DAG. They correspond to the notation of the simulation model in the main text.

Table 4.2: Legend for Figures 4-2, 4-3, and 4-4. The (observed and latent) variables and the (constant) parameters cover the elements of the model underpinning the assumed data-generating process.

Variable/constant	Type	Description
$q(s_i)$	observed	Attribute for sender i.
$q(s_{-i})$	observed	Attribute for sender other than i.
$\mathbf{a}_s$	observed	Vector of sender attributes.
$q(r_i)$	latent	Attribute for receiver j.
E_t	observed	Event sequence observed at time t.
$\xi(E_t, T_{1/2}, \mathbf{a}_s)$	observed	Homophily statistic. Captures the extent to which a sender node is guided by past behavior of other senders.
$P_{\text{similarity}}(s_i, r_j)$	latent	Similarity matrix with the row corresponding to sender s_i and column corresponding to receiver r_j.
$P_{\text{influence}}(s_i, r_j)$	latent	Influence matrix with the row corresponding to sender s_i and column corresponding to receiver r_j.
$P_{\text{decision}}(s_i, r_j)$	latent	Decision matrix with the row corresponding to sender s_i and column corresponding to receiver r_j.
$T_{1/2}$	parameter	Half-life parameter.
π	parameter	Mixing parameter value between similarity and influence for decision matrix.

There are two additional scope conditions of the causal identification of social influence considered noteworthy. The sender attributes must be uncorrelated with the hazard rate (other than through influence or similarity), and they must be exogenous. I will briefly summarize both conditions and their diagnosis, and detailed them further in their respective subsection using directed acyclic graphs.

The first condition states that the sender attribute under consideration must not influence the event sequence directly, i. e., other than through social influence or similarity. Consider an example in which countries can ratify environmental treaties independently as a function of their attributes, such as climate vulnerability or political rights, or as a function of social influence, for instance by learning from politically or physically similar countries (see the toy example in the next section). The shuffle test fails to distinguish reliably between similarity and social influence if countries with a higher vulnerability (e. g., because they have longer coast lines) or more political rights ratify environmental treaties faster and this effect is temporally structured in a similar way as the social influence effect. It is sufficient to demonstrate the absence of an empirical correlation between the sender attribute and time to ensure identifiability of social influence in the shuffle test (provided there is sufficient statistical power). In the example below, this is the case as several social influence model terms are shown to have corresponding sender main effects with effect sizes close to 0 and *p* values close to 1.0, including climate vulnerability and political rights.

The second condition states that sender covariates must be exogenous – they must not be affected temporally by the development of the event sequence. The simulations above considered only temporally constant sender attributes for reasons of simplicity. Yet sender covariates are allowed to change over time – but only as long as they are not a function of the unfolding behavior cascade. Complex feedback loops like these bear some resemblance with network-behavior coevolution (e. g., La Fond and Neville, 2010) but are a distinct problem. For example, consider the possibility that the cascade of past ratifications of environmental treaties by countries may systematically cause stronger or weaker political rights or longer or shorter coast lines, with a time horizon resembling the time horizon of the social influence effect. Causal pathways like these are somewhat uncommon. No statistical remedy exists for this confounding, but the required assumption of exogeneity can often be justified on theoretical grounds, similar to the exclusion restriction for instrumental variables.

4.4.1 Baseline scenario

Figure 4-2 shows the DAG for the simulations described in the main text. This will be called the baseline scenario, and will be expanded with possible confounders later.

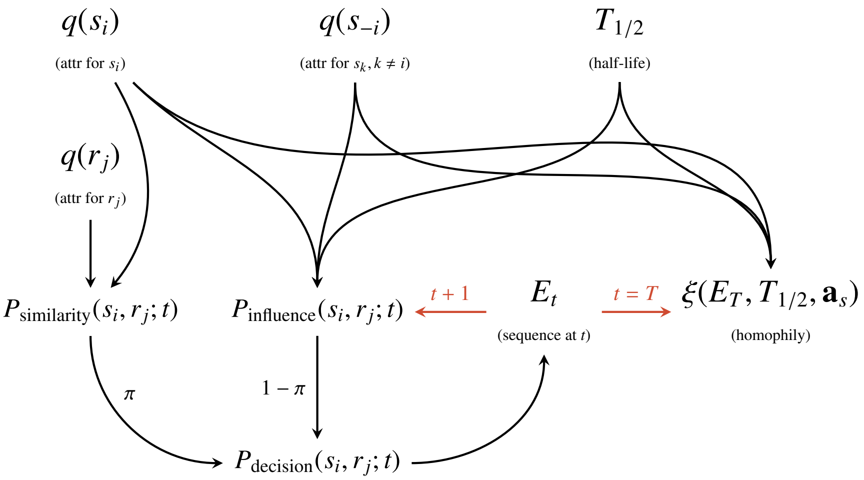

$q(s_i)$
(attr for s_i)
$q(s_{-i})$
(attr for $s_k, k \neq i$)
$T_{1/2}$
(half-life)
$q(r_j)$
(attr for r_j)
$P_{\text{similarity}}(s_i, r_j; t)$
$P_{\text{influence}}(s_i, r_j; t)$
$t+1$
E_t
(sequence at t)
$t = T$
$\xi(E_T, T_{1/2}, \mathbf{a}_s)$
(homophily)
π
$1 - \pi$
$P_{\text{decision}}(s_i, r_j; t)$

Figure 4-2: Base scenario underpinning the simulations in the main text. Unless otherwise stated, arrows are considered at time t.

The attributes of the sender nodes, for example the political rights or vulnerability of the sender country, and of the receiver nodes, for example treaties, are exogenous to the causal process, i. e., there are no cyclic paths or confounders leading to them, neither from within the diagram nor outside of the diagram. In empirical studies, the sender attributes are observed while the receiver attributes are unobserved.

Jointly, the attributes determine the probability that a tie between a sender and a receiver occurs at the next time step, through either of two processes (similarity or influence) or a blend between them as determined by the π parameter. In the simulations, π can be set to generate custom blends of the two scenarios. In empirical studies, this parameter is unknown. The probability in the similarity scenario is determined by the attributes of the sender and the receiver for each sender–

receiver dyad. Conversely, the probability for a sender–receiver dyad in the influence scenario is additionally affected by the attributes of the other senders and their past observed behavior in E_t.

Over time the dyadic decisions create the event sequence E_t, which both feeds back into new decisions (if influence plays a role) and lets us measure the extent of homophily in the overall sequence (where homophily can be comprised of similarity and influence as underlying causal mechanisms). The dependence of the influence probability on time through the path from E_t and the non-dependence of the similarity probability on time through the absence of a temporal path from E_t is what enables the shuffle test to discriminate between the two mechanisms in empirical studies using the relational event model by breaking the temporal order.

(Note that the temporal, orange edges are not cyclic because they are not simultaneous; I use an abridged DAG notation because otherwise $3t$ additional nodes in the diagram would have to created.)

4.4.2 Extension: attribute influence on event sequence

One of the assumptions of the model is that the sender attributes do not have a direct influence on the event sequence. To examine this as a scope condition of the shuffle test, Figure 4-3 adds such causal edges to the DAG.

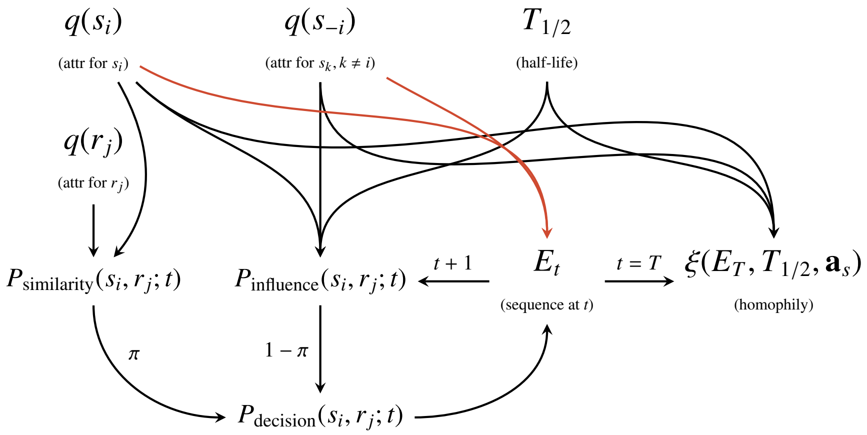

$q(s_i)$
(attr for s_i)
$q(s_{-i})$
(attr for $s_k, k \neq i$)
$T_{1/2}$
(half-life)
$q(r_j)$
(attr for r_j)
$P_{\text{similarity}}(s_i, r_j; t)$
$P_{\text{influence}}(s_i, r_j; t)$
$t+1$
E_t
(sequence at t)
$t = T$
$\xi(E_T, T_{1/2}, \mathbf{a}_s)$
(homophily)
π
$1 - \pi$
$P_{\text{decision}}(s_i, r_j; t)$

Figure 4-3: A possible source of unidentifiability of social influence: Sender attribute influence on the event sequence.

A causal effect of sender attributes on the event sequence means in an empirical context that an attribute is correlated with the hazard rate. In my running example of treaty ratification, this could manifest itself, for example, in a correlation between countries' climate vulnerability and their timing of ratifying environmental treaties – i. e., states that are more at risk because they have longer coast lines ratify faster, independently of how other countries behave.

Such a correlation between the sender attribute and the event sequence would make social influence unidentifiable because causality would bypass the decision probability. The shuffle test permutes the temporal sequence and would therefore shuffle both the temporally structured influence effect and the temporally structured sender attribute effect simultaneously to create a new null distribution – hence identification would fail.

Given the other scope conditions of the shuffle test, it is consequently sufficient to demonstrate the absence of causal paths between the sender attributes and the event sequence to ensure identifiability of influence in the shuffle test. Practically, this can be done simply by showing that there is no correlation between the sender attribute and time. For example, in the toy analysis in Section 4.7 I find small correlations of 0.13 or 0.11 between climate vulnerability or political rights and time, with *p* values for the correlations of close to 1.0, which translates into conditional main effects for climate vulnerability and political rights in the model of around 0.

In empirical applications of the shuffle test, the recommended procedure is to theorize about possible correlations between the sender attributes of interest and time. If they seem plausible, I recommend ruling these correlations out empirically by testing the correlation between the sender attributes in question and time. This test can be unconditional because the goal is to rule out the possibility of a causal path from the sender attributes to the event sequence, irrespective of other nodes in the diagram. If a correlation between sender attribute and time is present, causal identification of social influence is increasingly jeopardized the stronger the correlation.

4.4.3 Extension: endogenous sender attributes

Often, sender attributes do not change over time (i. e., they are fixed, or constant) and exogenous (i. e., they are not affected by other variables in the DAG). For example, the length of a country's coast lines over its area, which measures climate vulnerability, can be reasonably assumed to be exogenous and fixed.

Sometimes, sender attributes can vary over time. As per Figure 4-2, time variability of sender covariates does not change the identifiability of social influence as long as the sender covariates are exogenous. For example, political rights can change over time with regime changes and other exogenous events.

However, social influence is no longer identifiable if the variability in sender covariates over time is endogenously affected by the event sequence. Figure 4-4 shows an aggravated version of a correlation between sender attributes and the event sequence, where the event sequence causally affects the sender attributes. The causation can be bidirectional over time. (As a reminder, orange arrows are temporal and therefore exempt from the acyclicality property of the graph in the abridged notation.)

The presence or absence of early behavior leads to increased or decreased sender attributes in this scenario. For example, if countries' treaty ratifications led to increased political rights or decreased climate vulnerability in the short term, the decision probability of countries with early ratifications would go up or down and lead to more or less behavior, respectively, in turn. In the physics of complex systems, this is known as a feedback loop.

Empirically, this possible source of non-identifiability is hard to detect by assessing correlations, as in the previous case, because the causal paths can now flow in both directions. The consequence of this endogeneity is that social influence indirectly affects the similarity-based probability: Causality flows from the influence probability to the decision, on to the event sequence, on to the sender attributes, on from a focal sender's attribute to the similarity probability, and from there back into the decision matrix at the next time step. The shuffle test then fails to distinguish between the similarity- and influence-based mechanisms because both are temporally closely related.

With this limitation of the shuffle test in mind, I believe this possible source of non-identifiability will affect only a small minority of empirical contexts because it happens only in the subset of cases where current events affect future sender attributes among the subset of cases where sender attributes are temporally variable to begin with. Theoretical considerations can help to rule out this confounding: Is the sender attribute temporally variable? If no, identification is unaffected. If yes, is it theoretically plausible that a sender changes its attributes as a consequence of its own or others' behavior? If no, identification is again unaffected. The assumption is similar to the exclusion restriction in instrumental variables regression and must be justified in a similar way on a theoretical basis.

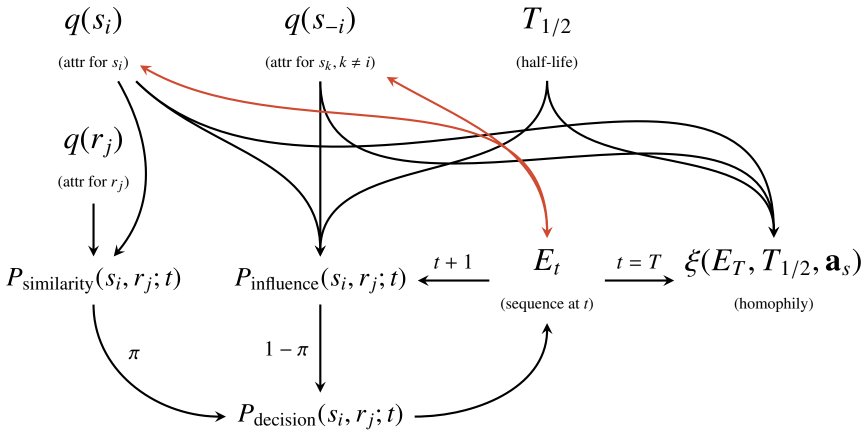

$q(s_i)$
(attr for s_i)
$q(s_{-i})$
(attr for $s_k, k \neq i$)
$T_{1/2}$
(half-life)
$q(r_j)$
(attr for r_j)
$P_{\text{similarity}}(s_i, r_j; t)$
$P_{\text{influence}}(s_i, r_j; t)$
$t+1$
E_t
(sequence at t)
$t = T$
$\xi(E_T, T_{1/2}, \mathbf{a}_s)$
(homophily)
π
$1 - \pi$
$P_{\text{decision}}(s_i, r_j; t)$

Figure 4-4: Feedback loops from the event sequence to the sender attributes.

Endogenous sender attributes make the relation between behavior cascades and the literature on network–behavior co-evolution (e. g. La Fond and Neville, 2010) apparent. In this literature, individuals' current network ties can influence their future behavior, and their current behavior can influence whom they choose to connect with in the future state of the network. In my model of behavior cascades, the probability of sender s_i to choose receiver r_j in the influence regime $P_{\text{influence}}(s_i, r_j; t)$ is a function of the similarity between sender s_i and the other senders, $1 - |q(s_{e^*}) - q(s_i)|$ (see Equation 4.5). This dynamically changing similarity matrix can be interpreted as a network matrix similar to the network in a co-evolutionary model, except that it is the result of two attributes rather than a social tie and it is based on the same variable that also defines the behavior. The similarity-based probability $P_{\text{similarity}}(s_i, r_j; t)$, on the other hand, is not the immediate result of any network matrix involving the other senders. It would correspond to the behavior variable in a co-evolutionary model. In contrast to the baseline scenario, the behavior cascade model with dynamic causality from the event sequence to the sender attributes as depicted in Figure 4-4 can therefore be interpreted as a special case of the co-evolutionary model: The attribute-based behavior affects the attribute similarity network underpinning the influence effect through changes in the event sequence, and the attribute similarity network underpinning the influence probability in turn affects the behavior dynamically through changes in the attributes as a consequence of the updated event sequence. The link between the behavior cascade model presented here and the co-evolutionary perspective is hence rather weak and holds only in special cases, but a relation exists.

4.5 Hypothesis testing with randomisation inference

In this section I explain how to use randomisation inference to hypothesis testing in the context of relational event models and bipartite event cascades. I am interested in distinguishing specific similarity variables that cause either independent similar behavior or act as influence channels. For example, if two countries both have long coast lines, does one country adopt marine policies by observing the policy adoptions by the other country (influence, policy diffusion), or does the similar attribute of having long coast lines cause both countries independently to adopt marine policies (similarity, confounding)? The attribute, in this case having long coast lines, is the explicit influence or correlation channel that is tested.

I use the fact that the relational event model employs temporal decay in forming the homophily statistic (the halflife parameter, as set out in Equation 2.20 in Section 2.4). The next section shows how temporal decay matters for the efficacy of the approach.

My approach can operate on multiple kinds of behavior, which are modeled as second-mode nodes (e. g., multiple different marine policies in the previous example), and multiple influence channels (e. g., coast lines and the strength of the fishing industry in the previous example), where each separate channel relates to all behaviors. Malang et al. (2019) showed an empirical example in which senders were parliaments, receivers were different legislative proposals, and influence channels were ideological compatibility, similar population size, similar location, and other possible channels. Previous approaches assumed that there was only one behavior and that the channels/-sources of influence or confounding were unknown.

In addition to this, my approach does not explicitly measure the social network ties but operates on bipartite sender–receiver event sequences, where the homophily statistic serves to uncover the variables or network relations that trigger influence (or similar behavior).

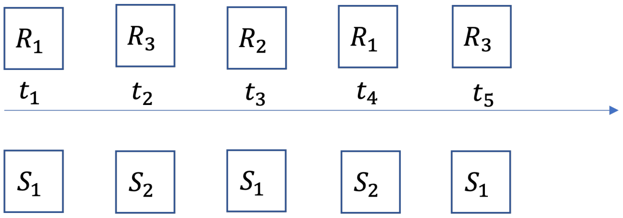


Figure 4-5: Diagram of a shuffled event sequence, showing a different ordering to that of the original, as an example of how the shuffling is done to test for the presence of influence.

The shuffle test described on this section follows the four steps introduced in Section 2.5. Figure 4-5 shows a possible shuffling of the event sequence introduced in Figure 4-1. In particular for bipartite relational event sequences, where we are shuffling on the order in which events took place. Step by step this is:

1. Estimate a bipartite REM, including a parameter for at least one homophily (ξ) statistic, $\hat{\theta}$. Observe one of two outcomes:

   (a) If the original homophily coefficient is not considered to be statistically significant at the 95% level in the estimation of the parameter relevant to ξ, we fail to reject the null hypothesis that there was neither influence nor similarity.

   (b) However, if the estimation of the model described in step 1 results in statistical significance for ξ, then the null hypothesis that there was neither influence nor similarity is rejected. As I showed in Table 4.1, the significance of the statistic could be caused by either influence or similarity. Proceed to Step (2).

2. Generate many (e. g., $k = 1,000$) synthetic event sequences $\widetilde{E}$ by randomly reassigning the time stamps t of the events e_t in E.

3. Estimate the same model as in Step (1) for each synthetic event sequence and save the ξ estimates $\widetilde{\theta}$ as a new empirical null distribution $\widetilde{\Theta}$.

4. Locate the original $\hat{\theta}$ in $\widetilde{\Theta}$. Observe one of two outcomes:

   (a) If $\hat{\theta} \leq \widetilde{\Theta}_{1-\alpha}$: Failed to reject the null hypothesis that $\hat{\theta}$ represented the effect of prior similarity (rather than influence).

   (b) If $\hat{\theta} > \widetilde{\Theta}_{1-\alpha}$: Null hypothesis that $\hat{\theta}$ represented the effect of prior similarity (rather than influence) is rejected.

This shuffle test is illustrated in Figure 4-6 for the two models reported in Table 4.1. Indeed, Model 1 with $\pi = 0.1$ (influence) yields a $\hat{\theta} > \widetilde{\Theta}_{0.95}$ while Model 2 with $\pi = 0.9$ (similarity) yields $\hat{\theta} \leq \widetilde{\Theta}_{0.95}$, as expected.

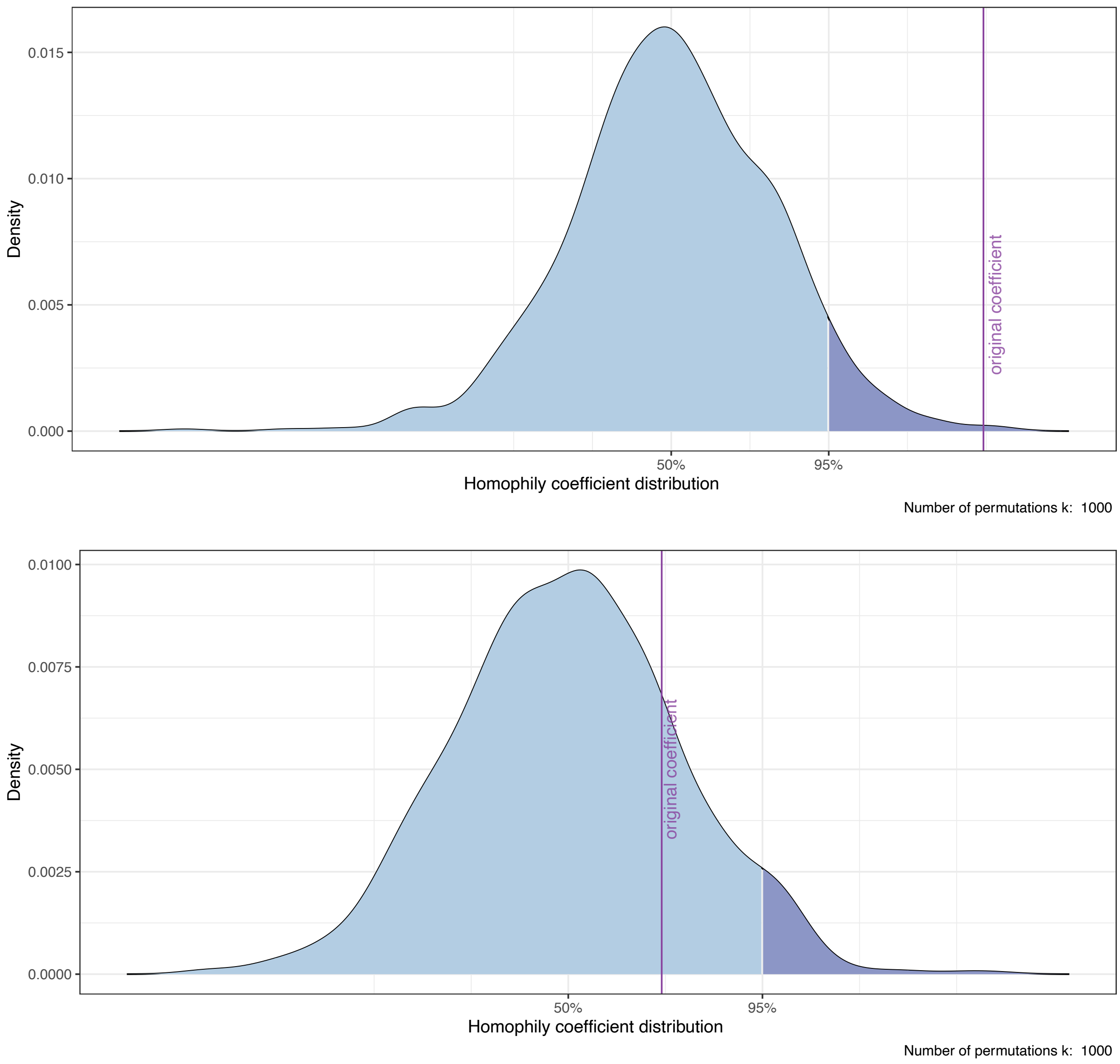
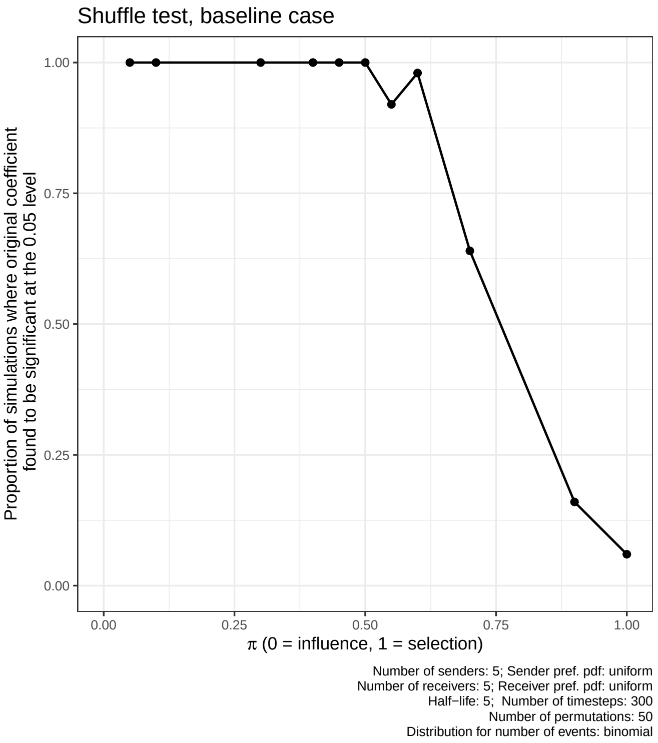


Figure 4-6: Application of the shuffle test to Model 1 (top) and Model 2 (bottom) from Table 4.1.

Having described the mechanics of the approach, I will now turn my attention to the scope conditions under which the shuffle test is effective and the conditions under which it breaks down, using Monte Carlo simulations. I will answer two questions: (i) Up to what level of π can social influence be reliably distinguished from selection if both are present in the same data-generating process? (ii) Under what other parameter conditions of m, n, T, and $T_{1/2}$ does the shuffle test break down and lose its discriminatory power?



Number of senders: 5; Sender pref. pdf: uniform
Number of receivers: 5; Receiver pref. pdf: uniform
Half−life: 5; Number of timesteps: 300
Number of permutations: 50
Distribution for number of events: binomial

Figure 4-7: Simulation results for mixtures of influence and selection in the data-generating process under plausible conditions.

4.6 Scope conditions of the shuffle test

I established in the previous section that the shuffle test can distinguish social influence from prior similarity when the data-generating process is a mixture predominantly dominated by influence or by similarity (e. g., $\pi = 0$ or $\pi = 1$). However, some empirical applications may feature a mixture of similarity and influence in their underlying data-generating process.

For instance, if the senders are states, the receivers are policies, the ties are instances of adoption of policies by the states, and the attribute is a dummy variable indicating whether the state is red or blue, then two mechanisms may compete: Similarity may dictate that states of the same color adopt the same policies at a similar rate, and influence may dictate that states adopt policies that were recently adopted by other states of the same color ("policy diffusion"). It is conceivable that both mechanisms occur in parallel (e. g., $\pi = 0.3$ or $\pi = 0.6$). Can I identify the share of influence under a given π regime?

Figure 4-7 extends the simulations from Figure 4-6 to different π regimes. The simulation results are based on $n = 5$ receivers, $T = 300$ time points, and a half-life parameter of $T_{1/2} = 5$, like in the previous simulations as a baseline case. The horizontal axis shows different values of π from 0.0 to 1.0. For $\pi \in \{0.05, 0.1, 0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7, 0.9, 1.0\}$, $\gamma = 100$ simulations were run with $k = 50$ permutations per simulation. The vertical axis shows the proportion of the 100 simulations in which $\hat{\theta}$ was in the right tail of $\widetilde{\Theta}$ with $\alpha = 0.05$ (one-sided test). The curve shows the results for each π regime: The higher the curve on the y-axis, the larger the probability that influence is detected in the relational event sequence through obtaining a significant result with the shuffle test.

For values of $0 < \pi \leq 0.5$ the test is highly sensitive to the presence of influence. The more influence in the data-generating process, the more significant results the shuffle test produces. There is one caveat. Cases where there is only selection but no influence effect ($\pi = 1$) yield slightly too many significant results (false positives). I do not consider cases where the data-generating process is purely driven by influence ($\pi = 0$), since in a purely social influenced world, where agents exclusively follow each other's behaviour, the simulated data is extremely dependent on the starting conditions (i.e. the attributes of the senders and the receivers), and does not yield consistent results.

Compelling evidence for social influence can thus only be produced in empirical settings where the data-generating process is dominated by social influence. If similarity plays a strong role, the homophily coefficient will be significant, but likely not the shuffle test. This has implications for how the shuffle test must be interpreted: It is a technique to assess the weight of evidence of influence versus similarity; it should not be interpreted as a tool for detecting small traces of social influence in a data-generating process that is otherwise dominated by similarity – unless one increases the statistical power of the test, as shown below.

4.6.1 Further scope conditions of the shuffle test

Having established that the shuffle test works in the baseline case (Section 4.6), I will examine further scope conditions of the test here. The discriminatory performance of the test breaks down when (i) there are few time steps (i. e., fewer than about 300 in the simulations); (ii) the half-life parameter is too large, which means there is little temporal decay in the homophily statistic; and (iii) the number of receiver nodes is very small (i. e., smaller than 5 in the simulations). Violations of these scope conditions decrease the statistical power of the test. Combinations of violations aggravate the lack of statistical power while a violation in one area can be compensated for in another area. The more favourable each condition, the smaller the traces of influence in the data that can be identified (i. e., the higher the sensitivity of the test). The lack of a significant result indicates either exclusive similarity, as opposed to influence, or a lack of statistical power because of a short observation period, few events per time span, a slow decay of homophily, or a small number of behavior alternatives that can spread interdependently. The presence of a significant result, in contrast, conclusively demonstrates social influence.

Unequal event frequencies over time, such as bundling of behavior towards the end of the observation period of a cascade because of a deadline, are no cause for concern because the temporal permutations retain the temporal frequency distribution of events in the observed cascade.

Like in any regression model, the usual assumptions hold. In particular, there must be no omitted variables; all confounders must be controlled for.

Figure 4-8 shows how the performance of the shuffle test is altered by the number of time steps (left panel) and the half-life parameter (right panel). A longer half-life and a shorter event sequence both undermine the statistical power of the test. The shuffle test breaks down with short cascades (or, equivalently, a small event rate); fewer observations lead to less statistical power like in any hypothesis test.

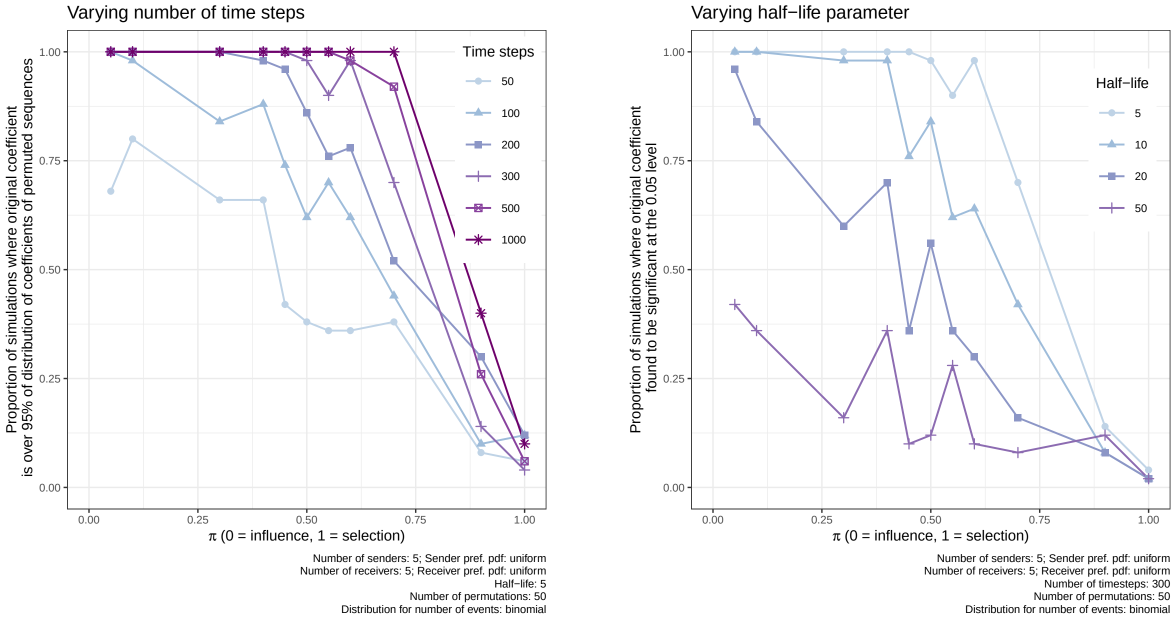


Figure 4-8: The shuffle test breaks down with a large half-life parameter or few time steps.

Additionally, a longer half-life parameter in the population process leads to long-term dependence, which increasingly resembles an atemporal similarity effect, diminishing the discriminatory power of time and rendering the shuffle test ineffective (cf. Figure 2-2). This means the shuffle test is only able to detect somewhat short- to medium-range influence. As reported by Malang et al. (2019) in an empirical case, the choice of the half-life parameter during estimation matters relatively little with a fixed data-generating process; it is the population half-life parameter that matters for test performance. The size of $T_{1/2}$ in the population at which the test breaks down is a function of the event rate and should not be evaluated in absolute terms. I expect that longer cascades should make up for longer half-life in empirical settings. On a positive note, if the shuffle test detects social influence, this is strong evidence for contagion.

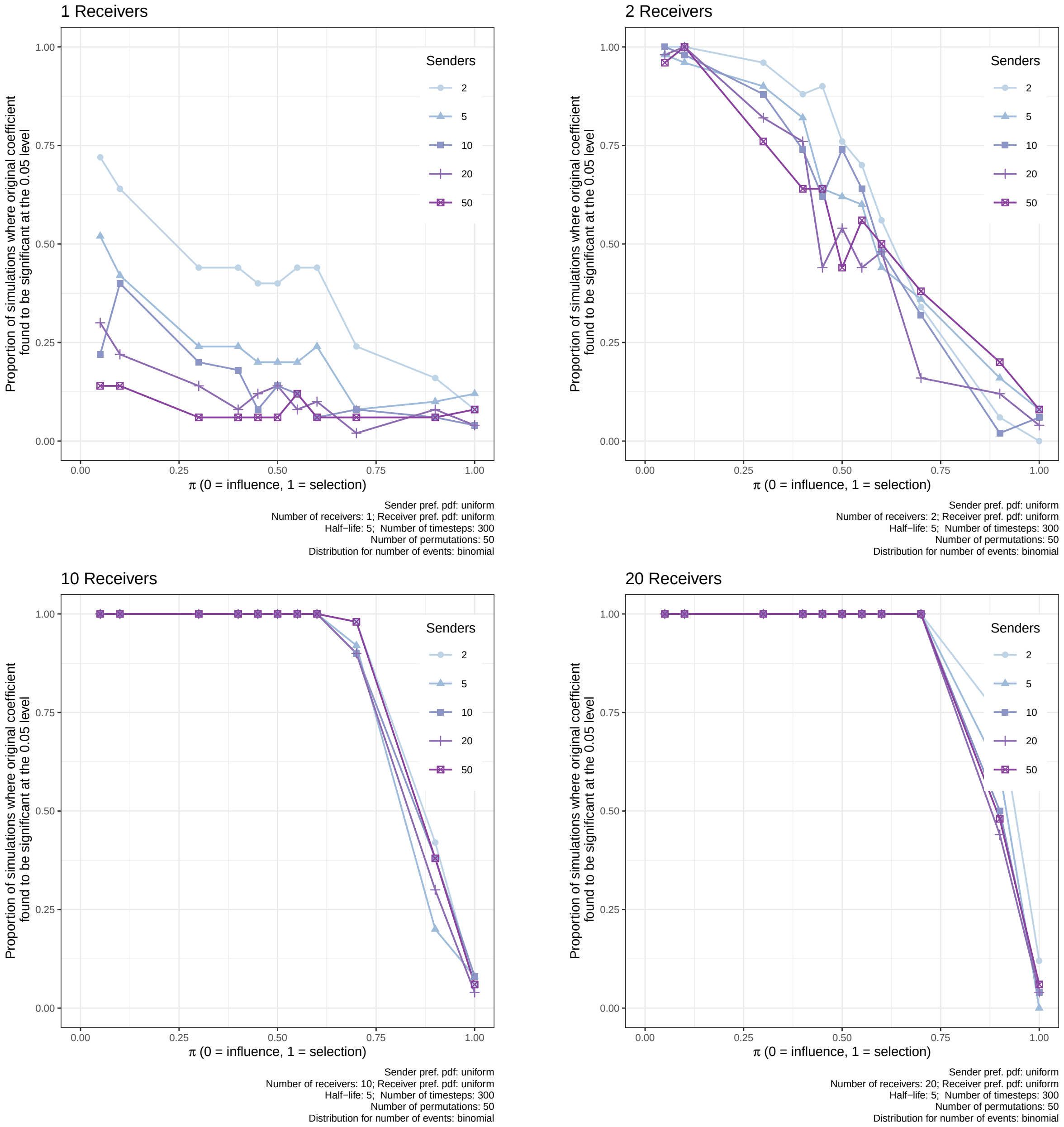


Figure 4-9: Simulation results for mixtures of influence and selection in the data-generating process when parameters *m*, and *n* are changed.

Figure 4-9 shows different combinations of the number of senders and receivers, with the remaining parameters duplicating the baseline configuration. Fewer senders (actors) and more receivers (behavior types) lead to better test performance. The performance breaks down with one or two receivers, e.g., the case of a single treaty that can be ratified successively by countries. This result can be interpreted as a lack of statistical power as there is little evidence for influence. The discriminatory power becomes worse if this one behavior type is successively displayed by many senders – again a case of little statistical power. With the same m and n, the problem would increasingly disappear with longer cascades of repeatable behavior. Conversely, with only two senders and many receivers, increasingly smaller traces of social influence can be identified.

4.7 Example: ratification of environmental treaties

I now apply the procedure to an empirical example. The literature on international environmental politics posits that interdependencies between countries explain the ratification of international treaties (e. g., Simmons, 2000; Elkins et al., 2006; Bernauer et al., 2010; Perrin and Bernauer, 2010; Hugh-Jones et al., 2018; Shim and Shin, 2020). Political rights and civil liberties (e. g., as measured in the Freedom House, 2020 Index) have been shown to play a role in the timing and magnitude of environmental treaty ratification at the country level. Conclusive evidence on whether countries are influenced by countries with similar values on these variables is missing; positive results have been produced at the country level, assuming independence. Similarly, GPD per capita and climate vulnerability have been shown to increase the speed of ratification, and the past number of ratifications in the same geographical region has been shown to be positively associated with ratification decisions – a homophily effect (Bernauer et al., 2010).

I draw inspiration from this research and examine for a set of eleven environmental treaties (those with at least 170 ratifications)[1] whether homophily plays a role in the hazard of treaty ratification – regarding political rights (Freedom House, 2020), climate vulnerability (Chen et al., 2015), similar population size, and shared borders.[2] For each homophily variable I examine whether countries with similar levels on these variables either ratify (or abstain from) the same treaties independently (similarity) or whether they influence each other's ratification behavior (social influence). This is a much-reduced toy dataset and serves only the purpose of illustration.

[1] Basel Convention; Convention on Biodiversity; Cartagena Protocol; CITES; Kyoto Protocol; Paris Agreement; Ramsar Convention; Stockholm Convention; UNCCD; UNFCCC; Vienna Convention.

[2] Downloaded from `https://github.com/geodatasource/country-borders` (accessed 28 February 2021).

Across the 189 countries and eleven treaties, there are 1,971 ratifications and 670,741 null events (i. e., country–treaty–day combinations without ratifications), spread over 1,456 strata in the counting process dataset, with a start date of 14 January 1974 and an end date of 21 August 2020 (the day the data analysis concluded). The first 74 events were discarded as burn-in for the formation of model terms, leading to an actual start date in 1980. The political rights, climate vulnerability, and population observations are time-varying covariates; geographical neighborhood is time-constant. Political rights and civil liberties are collinear, hence only political rights were used. The results are very similar if civil liberties are modeled instead of political rights.

Figure 4-10 shows the expected increase in the log hazard ratio – the log odds that a ratification event occurs between a sender country and a receiver treaty given that it has not occurred yet – when the independent variable on the *y* axis is increased by one unit, all else being equal. 95 percent confidence intervals are drawn around the point estimates. The first four rows show the homophily terms. The next three rows show monadic main effects for the same variables (except the neighborhood term because it is intrinsically relational). The final three rows show monadic control variables: GDP per capita in constant 2010 U.S. dollars (using World Bank data); the sender country's recent number of ratifications (with each past event geometrically weighted with a half-life parameter of 50 days as per Equation 2.20 like all other model terms); and the recent number of ratifications of the receiver treaty.

The homophily terms show positive, significant results. Interpretation of effect magnitudes is being complicated by the geometric weighting with the half-life parameter and the standardization in the homophily statistic, but one could generate predicted probabilities conditional on different timing of past events if the effect size were of interest.

Figure 4-11 shows the shuffle test for the four homophily terms. In all four cases timing matters, and the null hypothesis that there is no social influence at work can be rejected. States make ratification decisions not merely independently at similar times as other states with similar political rights, climate vulnerability, population size, and geographical location, but display a higher risk

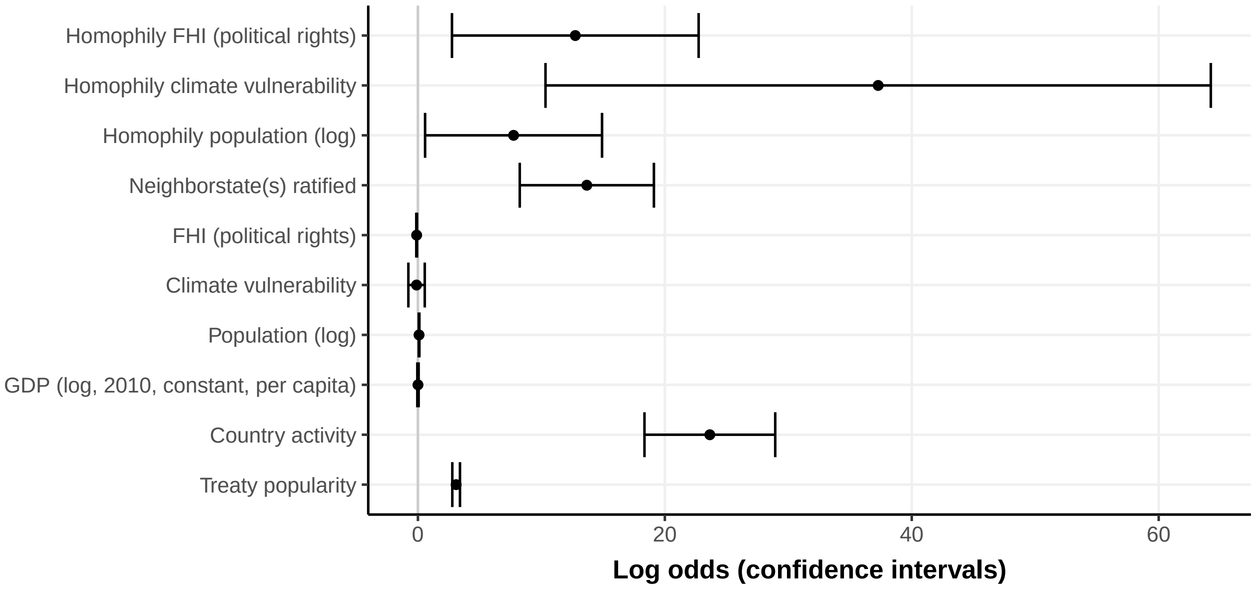


Figure 4-10: REM results on treaty ratification.

of ratification when politically, geographically, and demographically compatible countries ratified the treaties in question in the recent past.

However, this toy example serves to illustrate not only the viability of this approach, but also to discuss potential limitations: First, three of the four variables used for homophily and the shuffle test were time-varying, with only geography being time-constant. In the simulations I only considered time-constant attributes for reasons of simplicity. The results theoretically hold for time-varying attributes as well. In Section 4.4.3 I showed using directed acyclic graphs under which situations identification problems can be expected with time-varying sender attributes. In theory, the null distribution maintains the temporal distribution of ratifications, which means exogenous shocks on both the attribute and the ratification behavior are already factored into the shuffle test. There is a residual risk that the test is less effective if all probability mass is concentrated on very narrow time periods. But then the homophily terms would not display a significant result in the first place, hence this should not diminish the efficacy of the test. Future research should evaluate the connection between identifiability and exogenous shocks on the system more carefully when the attribute variables vary over time, but I do not see any immediate cause for concern.

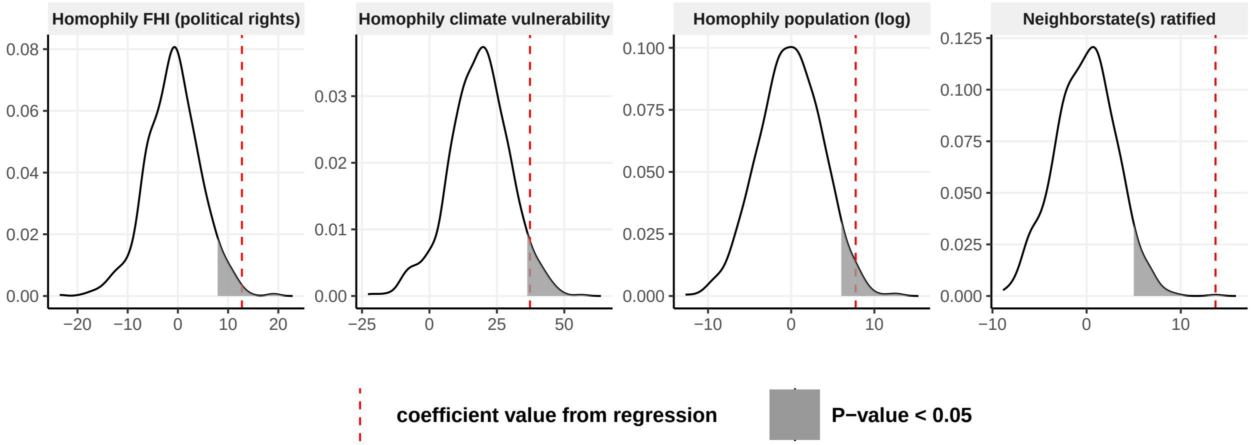


Figure 4-11: Empirical distributions and original homophily coefficients on all four homophily terms considered: political rights, climate vulnerability, population and neighbouring state.

Second, just like in other regression models, we need to be wary of omitted confounders. Here, I did not control for trade as a possible homophily term[3] or other variables that may explain both ratification behavior and incentives for social influence through similar political rights. I did control for the main effects as possible sources of temporal structure in the homophily effects, and they were conditionally uncorrelated with the hazard rate. The usual assumptions of likelihood-based inference apply. Theoretically, I do not expect a country's treaty ratification to exhibit short-term impacts on vulnerability and political rights because possible changes in these variables would likely have longer time horizons and would only be affected by environmental treaties if the treaties enabled significant environmental change or preservation. Hence reverse causality can be ruled out on theoretical grounds as per the conditions formulated in the previous section.

Third, the example was in good keeping with the scope conditions of the simulation results: $m = 189$ senders, $n = 11$ receivers, an assumed half-life parameter of $T_{1/2} = 50$ (though the true parameter in the population process is unknown), and sufficient statistical power with several thousand time steps and $1,971$ events. The simulations in Section 4.6.1 showed how the test would break down with fewer receivers – the most critical scope condition for the efficacy of the test. Indeed, if I arbitrarily pick four out of the eleven treaties and drop the remaining ones

[3] Strictly speaking $1-|q(s)-q(r)|$ would need to be replaced by the trade volume in Equations 2.21 and 4.5 because a network relation is used instead of an attribute.

from the analysis, the shuffle test no longer indicates significant social influence; for example the population homophily term drops out of significance in the shuffle test but retains its significance in the original model (not visualized here for brevity).

4.8 Conclusion

This chapter made the case for re-interpreting a wide range of categorical behavior across subfields as bipartite relational event cascades. Political science in particular will benefit from the application of corresponding methods by improving the scope of theories that can be developed and tested. We should start collecting observations on the behavior of states in the international system, legislators in parliaments, interest groups in policy processes or lobbying coalitions, and other actors operating in group contexts at a higher temporal granularity than the current practice of collecting annual observations. Doing so will permit better causal identification of endogenous theory because it is often the order of events that matters for explaining the choices actors make.

I developed a conceptual framework for reasoning about similarity and social influence in bipartite relational event models of actors and categorical behaviors. This includes an approach for simulating artificial event sequences with given relative levels of similarity and influence. Using these simulations, this chapter demonstrated how similarity and influence are confounded in empirical applications of the relational event model with a homophily statistic. As such, I introduced a shuffle test with the potential to rectify the identification issue and to test for the prevalence of influence as opposed to prior similarity. Simulations evaluated the sensitivity of this shuffle test to different theoretical conditions which may be found in plausible empirical settings. An empirical toy example further illustrated the method and its limitations.

The simulation results show interesting variation: The shuffle test can robustly identify influence when the data-generating process is overwhelmingly dominated by influence. It can identify a preponderance of influence over similarity, and its efficacy gradually declines, subject to different conditions, with increasing importance of similarity in the data-generating process. The implication is that a non-significant test result cannot be interpreted as an absence of influence. In line with the general thrust of statistical testing, only a significant result can lead to a rejection of the null hypothesis of the absence of social influence. A non-significant result, in contrast, may indicate either a preponderance of similarity or a lack of statistical power of the test through one of the ancillary conditions. These can be, among others, a small number of time points, a low density of events in the sequence (which is equivalent to a short sequence), a long time horizon of the influence, a large number of senders, or a low number of receivers/behaviors. The extent to which the test is affected by these conditions varies by empirical context. As these conditions all affect statistical power, each condition may be compensated for to some degree by improving another condition.

The shuffle test is easy to implement for any bipartite relational event model. It just requires randomization of the time stamps in the event sequence and re-estimation of the REM many times. The added benefit is that under many plausible real-world conditions the shuffle test permits identification of social influence. This is an important step as many theories posit social influence but lack empirical identification strategies. In some instances, this lack of identification techniques has led to an overselling of findings as social influence when they may well have been caused by prior similarity (termed the "spread of evidence-poor medicine via flawed social network analysis" in one such instance by Lyons, 2011).

The technique is applicable in situations with the following characteristics: (i) availability of fine-grained temporal observations about actors (Mode I) without much simultaneity in the observations; (ii) one or more kinds of binary behavior the actors can adopt once or repeatedly (Mode II); (iii) at least partial observability of each other's behavior by actors (and this may be restricted by attribute variables); (iv) observability of one or more time-invariant (or potentially time-varying) attributes (or network relations) that presumably structure the behavior in a relational way ("channels" or "sources of influence"); (v) no omitted variables (including exogeneity of the attributes) and no correlation of the attribute with the timing of behavior (other than through similarity or influence); (vi) sufficient statistical power.

As per condition (iv), the procedure is applicable in cases where there are multiple homophily terms with different influence channels. However, future research will need to evaluate the robustness of the test to such increases in model complexity; they may need to be offset by other sources of statistical power. In particular, future research should evaluate possible consequences of correlations larger than zero between multiple influence channels. Relatedly, a strength of the technique is that it can use network relations in lieu of attribute values. Malang et al. (2019) demonstrated in an empirical application on inter-parliamentary diffusion of subsidiarity concerns about legislative proposals by the European Commission how the shuffle test can be applied to network relations like geographical contiguity, in addition to attribute-based homophily, for example based on shared ideology.

The shuffle test assumes that there is no omitted confounding variable that explains both the attribute (or the resulting similarity) and the adoption of behavior (Xu, 2020), as is standard in regression modeling (condition v). It furthermore assumes that there is no contagion occurring within time points (hence the word "fine-grained" in the first condition) and, by virtue of the event history roots of the relational event model, that there is no simultaneity in decision events, as per Equation 2.19, Section 2.4 (Xu, 2020; Butts, 2008). While this may seem like a restrictive assumption, several methods (e. g., Efron, 1977) have been tried and tested to break tied events ("interval censoring"). The default method used in `R`'s `survival` package is an exact test that considers all possible combinations of temporal orders in the likelihood function. Bipartite relational event modeling with a shuffle test is somewhat computationally involved; while the randomization in the shuffle test can be parallelized, the computation of the homophily statistic in Equation 2.21 requires nested iterations over the past event sequence and thus leads to increased computation time. So far, we have applied the procedure to sequences with several thousand events and several hundred nodes without problems on a contemporary multicore CPU. Lerner and Lomi (2020) have recently suggested a sampling approach that may solve the computational burden, but its consequences for the shuffle test have not been evaluated yet.

In this exposition, I chose a simple REM specification for discrete time points using a stratified Cox model. I expect the results to hold with continuous-time REM specifications, but this should be evaluated in future research. An open question is also how to select the optimal half-life parameter. Empirical applications like the one reported by Malang et al. (2019) seem to be relatively robust to different parameter choices, but there is no conclusive evidence yet. The evidence shown here suggests that different half-life parameters in the data-generating process, rather than the empirical estimation, are consequential for the identifiability of influence, but this is different from the half-life parameter chosen for estimation.

I also expect that the approach is applicable no matter whether each behavior/receiver can be adopted multiple times per sender or only up to once. Such different scenarios can be accommodated in the model by specifying the risk set appropriately. Future research needs to verify the efficacy of the shuffle test in these different situations. It also needs to account for instances where the difference in the values of the attributes between senders is very large, and how this might play a significant role in the determination of presence of influence in the event sequence. In relation to the number of senders acting in an event sequence, and the total number of events, more research should be done to determine whether a large pool of possible actions by the senders might affect the way in which the shuffle test captures the presence of influence. With more options, influence might be harder to detect; alternatively, with fewer options the test might be over-sensitive to the presence of influence.

Furthermore, the present setup of the REM and shuffle test assume that the attribute variables do not change over time. While constant attributes should correspond to a range of plausible scenarios (e. g., geography, party family, or actor type), it is easy to imagine empirical applications where they are unrealistic. For example, policy diffusion may be affected by changing majorities in states over time, as per one of the examples mentioned above. In the absence of omitted variable confounding, I do not see any reason to expect time-varying attributes to invalidate the shuffle test, but future research should investigate more carefully.

Political science is currently moving towards better causal identification. Where experiments are not feasible, it is observational methods like the one presented here that can support causal identification. This is most difficult in endogenous settings. Analogously, experimental design in such settings is plagued by interference. Both with observational and experimental designs I therefore need to find ways to pinpoint, and isolate, social influence. This echoes Flache and colleagues' call that "more empirical work is needed testing and underpinning micro-level assumptions about social influence as well as macro-level predictions" (Flache et al., 2017).

Chapter 5

Bayesian co-estimation of network characteristics and node attributes with the exponential family of distributions

5.1 Exponential random network models

Estimation of parameters for network data using the exponential family of distributions, as described in Subsection 2.2.2, is based on comparing the observed network with all the possible configurations its nodes can make. An approximation has to be used for the cases in which the number of nodes in the network is more than 20, because the total number of configurations exceeds what is computationally feasible (Hunter and Handcock, 2006). Traditionally, ERG models are estimated using the methodology developed by Barndorff-Nielsen (1978). One of the most common software implementations of this estimation strategy in `R` is the `statnet` suite of packages. The estimates produced by this implementation are made with the different configurations of the network that only contemplate networks with different sets of edges, but have the same values for the node covariates.

For a network defined as $G = (Y, X)$ (as in Section 2.2.1), this means that all the possible edge configurations, $\mathcal{Y}$, assume the set of node attributes X to be the same. Those configurations where both the edges and some or all of the node covariates change are not considered under this traditional approach. The ERGM (Hunter and Handcock, 2006; Hunter et al., 2008) implementation makes the assumption that all the feasible edge configurations in $\mathcal{Y}$ are variations of the observed network, where only the connections between the existing nodes change. This is equivalent to saying that the data generating process where the observed and sampled networks come from, takes the values of the node covariates as a given.

Fellows and Handcock (2012) proposed a new estimation routine that considers a data generating process where viable samples are produced by toggling both the edges and the node attributes. This change expands the possible number of networks to be considered for inference to not just those in $\mathcal{Y}$, but rather $(\mathcal{Y}, \mathcal{X})$. Details on this toggling procedure will be presented in Section 5.3. This change increases the possible number of configurations from a large but countable number, to a very large one and possibly uncountable, depending on the kind of attribute to be toggled. For the purpose of this dissertation, we are only going to deal with binary attributes.

The estimation methodology developed by Fellows and Handcock (2012), however, suffers from the same chances of sampling degenerate networks in the approximation to the normalising constant as other methodologies that rely on maximum likelihood to estimate a network. As demonstrated by Handcock (2003), degeneracy in estimation occurs when the starting parameters are close to what is referred to as the boundary of the convex hull of the sufficient statistics. For the case of the exponential random graph model, Caimo and Friel (2011) developed a Bayesian estimation that uses an exchange algorithm to propose new values of the parameter, reducing the number of steps required to reach a stationary distribution, even when starting close to the degenerate region.

I now present the methodological foundations for Exponential Random Network Models (ERNM). As a motivating example, consider, three nodes, A, B, and C in an undirected network. A and B are assigned the red denomination, and C is assigned the green denomination. The number of possible connections with these three nodes is limited to $A-B$, $A-C$, $B-C$, $A-B-C$, $A-C-B$, $C-A-B$, the full network and the empty network. When we look into the assigned colours for the selected nodes, we lose information since the graphs $A-C$ and $B-C$ are considered to be the same graph. The Exponential Random Network Model (ERNM), as it was coined by Fellows and Handcock (2012), allows these two to be different graphs because $A-C$ and $B-C$ can emerge as two different graphs in the sampling process and in the approximation to the normalising constant, $c(\theta)$ (see Equation 2.3).

Another more illustrative example could be the analysis of a network of students and their smoking behaviour over time. In this scenario we are interested in determining whether the uptake of smoking by some individuals, i.e. the change of a binary attribute for one individual from A to B, happened before or after this individual developed ties to other individuals with attribute value B. When considering the attributes to be exogenous, as it happens in the temporal exponential random graph model, the possibility that there was a change of attribute that coevolved with the development of the network is lost. The coestimation routine introduced in this chapter is a first step in the development of Bayesian coestimating of parameters for network statistics and nodal attributes that uses the exponential random family of distributions.

Consider a network $G = (Y, X)$ as introduced in Subsection 2.2.1. Exponential Random Graph Models (ERGMs) look to calculate the probability of observing a particular configuration of nodes conditional on the node attributes:

$$p(Y = y|X = x, \theta) = \frac{e^{T(y,x)\theta}}{\sum\limits_{y' \in \mathcal{Y}} e^{T(y',x)\theta}}, \tag{5.1}$$

where $T(y, x)$ represents the vector of sufficient statistics that characterises G. The ERNM presented by Fellows and Handcock (2012) extends this model to

$$p(Y = y, X = x|\theta) = \frac{e^{T(y,x)\theta}}{\sum\limits_{(y',x') \in (\mathcal{Y},\mathcal{X})} e^{T(y',x')\theta}}, \tag{5.2}$$

suggesting that the space of possible networks accommodates ones with the same set of connections, y, but different node attribute values, x and x'. Alternatively,

$$p(X = x|Y = y, \theta) = \frac{e^{T(y,x)\theta}}{\sum\limits_{x' \in \mathcal{X}} e^{T(y,x')\theta}}, \tag{5.3}$$

represents a Gibbs/Markov field when the process satisfies the pairwise Markov property (that is, that when nodes i and j are not connected, then X_i and X_j are conditionally independent given all other node attributes).

The contribution of this chapter is a version of the sampler originally developed by Fellows and Handcock (2012) that allows for more control in the way networks are proposed, in the form of a user-defined probability of selecting to toggle from the existing edges in the network or all the possible dyads. I also modify the probability of selecting whether to toggle an element in Y and an element in X to a proportional probability based on the number of elements to toggle in X and the number of elements to toggle in Y. Additionally, I apply the estimation strategy from Caimo and Friel (2011) to produce a posterior distribution of parameters based on the observed network.

My motivation for developing this methodology is twofold: first, the move from a routine that produces estimates for $p(Y = y|X = x, \theta)$ to one that produces estimates for $p(Y = y, X = x|\theta)$ is needed to capture the nature of networks where node attributes and network connections occur at the same time. In particular, in the context of causal inference in the presence of interference, a more robust estimation routine like this one is needed to accommodate for the complex interdependencies present interventions where the treatment can spill over between individuals through their social connections. Second, the model developed in this dissertation is based on the endogenous relationship between network characteristics and node attributes in cross-sectional data. I pursue the development of a model for the co-evolution of dyad and nodal covariates, which can only be achieved when observing a network through time - the model presented here is the first step in that process. The Bayesian implementation ensures a more stable way to navigate the parameter space, as shown by Caimo and Friel (2011).

The rest of the paper is organised as follows. Section 5.2 explains the background and recent developments that deal with network estimation that considers nodal attributes in the estimation. Section 5.3 shows the methodology for sampling networks that consider nodal attributes, as well as the way parameters are estimated. Section 5.4 shows the estimation of an example on Sampson's monk data. Section 5.4.1 presents the results from a simple simulation study that shows the estimation methodology recovers the true values of a parameter used to generate random networks, and Section 5.5 concludes.

5.2 Recent developments stemming from the work of Fellows and Handcock (2012)

Since the publication of Fellows and Handcock (2012), there have been other approaches to consider the variability of node attributes in network sampling. Thiemichen et al. (2015) and

Thiemichen et al. (2016), use Caimo and Friel‘s methodology for parameter estimation to include nodal random effects. The traditional ERGM approach models all possible heterogeneity in the nodal attributes as included in the covariates, influencing the global structure of the network. However, this leaves possible heterogeneities in the nodal covariates unaccounted for. The proposed approach of Thiemichen et al. includes random, node specific statistics in a traditional ERGM to account for said heterogeneity. Krivitsky et al. (2009) presented a similar approach using a latent cluster random effects model to represent heterogeneity in the observed node covariates.

Fosdick and Hoff (2015) propose a methodology to test whether dependencies between nodal attributes and the network itself exist, and a joint attribute and network modelling approach for when it does. The joint model allows for estimation and inference on the dependence between and within the network and attributes. Additionally, it provides a way of handling and predicting missing network and attribute data. Li (2015) proposed a sampling algorithm similar to Fellows and Handcock (2012) for temporal ERG models, and extended it to estimate longitudinally observed networks. Most recently, Yan et al. (2019) develop a model that allows for node heterogeneity via node-specific parametrisation, and quantifies the extent of heterogeneity in terms of outgoingness and incomingness of each node by different parameters. To the best of my knowledge, this is the first time a model that considers the joint distribution of network connections and individual attributes using the Bayesian paradigm to estimate the vector of unknown parameters.

Stochastic Actor-Oriented Models The exponential random graph model, as it was presented in Section 2.2.2, is only one of the tools researchers can use to make inferential statements about the structure of a network. A methodology that has been developed in parallel is the Stochastic Actor-Oriented Model, or SAOM. Developed by Snijders (2001), this considers networks observed at different times, and regards them as snapshots of a continuously evolving process. The probability of change between the node connections are determined by the current state as in a Markov process. In between observations, the process evolves as follows: each actor individually determines whether to send a new dyad or remove an existing one, following an objective function that incorporates local and global information. Each one of these decisions are steps that occur one after another in the modelling of the observed network. At every one of these steps, there is one single change to evaluate. For every one, the model evaluate whether it increases the likelihood of observing the actual network and decide whether to keep that change or not accordingly.

The range of outcomes considered in the objective function can be extended to include multiple networks, or actor-level attributes. This last inclusion allows for an estimation procedure that considers co-evolving networks, where node attributes network characteristics are both considered as endogenous, in order to produce the observed network. This model provides a rich statistical structure for the analysis of social networks, and has seen many methodological developments since its inception (see Snijders et al., 2006, 2010; Snijders and Pickup, 2017, and Snijders et al., 2021 for the software implementation called RSiena), but it is not the focus of this work.

5.3 Algorithms

Two algorithms comprise the bulk of the presented methodology. As introduced in Section 2.3.2, following the methodology developed by Caimo and Friel (2011), I observe a network and look to find a vector of parameters θ that is compatible with the network observed. The first algorithm I am going to present in detail describes how to get the network samples, and the second one describes how to move through the parameter space to reach a valid sample of vectors of θ.

5.3.1 Sampling algorithm

The sampling algorithm follows the methodology presented by Fellows and Handcock (2012). The evolution of this algorithm started with Strauss and Ikeda (1990), and is in use in the current implementation of the ergm package (Krivitsky et al., 2020). Fellows and Handcock (2012) designed a version of this algorithm allowing for a key innovation: the toggling of certain attributes as well as dyads in the network. The algorithm presented here is a modified version of that algorithm to allow for a way to propose innovations to dyads or attributes proportionally to the total number of items to change. This innovation will be presented in detail in Algorithm 1.

I present a brief description of the steps for this algorithm. Consider $G = (Y, X)$, a network and T_G a vector of sufficient statistics based on our understanding of G; from G we know the number of nodes and the types of covariates in X. The choice of sufficient statistics is a key part of models for network inference. These represent how the analyst summarises the observed network. Some examples include the number of edges in the network or the sum of one of the attributes for all nodes i and j where $(i, j) \in Y$. For a more detailed explanation on the broad range of statistics see Morris et al. (2008). θ is a vector of parameters connected to these sufficient statistics. The goal of this algorithm is to produce a network that is conditional on the value of θ:

$$Y = y, X = x|\theta.$$

To begin, generate a random starting network $G_0 = (Y_0, X_0)$, composed of a random set of edges, Y_0, and a random vector of covariates for the nodes present in the network, X_0. Set l_1, the number of iterations, to be a sufficiently large to ensure convergence. For step $i = 1$, $G_i = G_0$, the initial random network. For every $i \geq 1$ the algorithm chooses, based on a specific rule (described in detail below), to toggle an element in the current network and create a proposed network. It then evaluates whether this proposal is a better fit to θ than the current one. The sampler developed by Hunter et al. (2008) produces a new network by choosing whether to toggle an edge that already exists in the network, or a random empty dyad, with a probability of 50% (Morris et al., 2008). After the ł$_1$ iterations, the algorithm will produce a network conditional on the selected value of θ^*.

The justification from Morris et al. (2008) is that if the chance of toggling an empty dyad and not an existing edge is proportional to the amount of dyads vs edges in the current network of the sampler, there would not be enough mixing of the McMC chain, making convergence slower. Only choosing from the set of dyads for the next toggle results in the sampler remaining in the same state for multiple steps of the Markov chain. The proposed innovation I present in this chapter is the ability to select with which probability dyads are selected when sampling the network, to be adjusted according to the characteristics of the network. This probability can be smaller than 0.5, since mixing of the McMC chain also occurs with the toggle of the node attributes. I refer to this probability p_{edge}, following Fellows and Handcock (2012) and Li (2015).

In addition to toggling dyads, the ERNM sampler implemented here also toggles nodal attributes. One alternative for this probability is to follow a proportional probability of toggling an element in Y or an element in X. In other words, the sampler chooses to toggle elements in Y with probability p_{dyad} or to toggle elements in X with probability $1 - p_{dyad}$, where

$$p_{dyad} = \frac{n * (n-1)}{n * (n-1) + k},$$

and k is the number of attributes available for toggling in X. A random number drawn from a uniform distribution between 0 and 1 will allow a proportional navigation of the possibilities of toggling a dyad or toggling an attribute. However, if this value is close to 1, i.e. if the number of attributes to toggle is relatively small in relation to the total number of dyads in the network, a user-defined probability can be used. Algorithm 1 has the detailed steps necessary to produce network samples conditional on θ.

In the case that p_1 (defined in Algorithm 1) is larger than p_{dyad}, the algorithm moves to toggle an element in X. The version of the toggle described in Algorithm 1 is designed for binary attributes. However, if the attribute to be toggled in X was continuous and bounded, the proposal X^* would be created by adding a random innovation within the expected bounds of X. In the next subsection I give an example of how the designed sampler works and how it compares to the ERGM sampler, as well as comparing what happens when we allow for toggles of the edges and the attributes, and when we allow only the edges to be toggled.

5.3.2 Estimation algorithm

This estimation algorithm follows the methodology developed by Caimo and Friel (2011) to arrive at a posterior distribution of θ based on the observed network, as described in Section 2.3.2. Step

Algorithm 1: Network sampling algorithm

```
Input: $n$, number of nodes; type of $X$, $\theta^*$, $T_G$, $p_{edge}$, $p_{dyad}$
Output: Network conditional on $\theta$ via $T_G$
Produce random network $G_0 = (Y_0, X_0)$ based on $n$ and type of $X$;
For a given number of iterations, $l_1$;
while $i < l_1$ do
    Draw two random numbers $p_1$ and $p_2$ from a uniform distribution between 0 and 1.
    if If $p_1 < p_{dyad}$: Toggle element in $Y$ then
        if If $p_2 < p_{edge}$: Toggle existing edge then
            Choose one of the existing edges at random and change it from an edge to a non-edge. This
              creates a proposed set of edges $Y^*$, and proposed network $G^* = (Y^*, X_i)$.
            We now have two networks, the current network $G_i$ and the proposed one $G^*$. We calculate the
              change statistics (as described in Section 2.2.2) between $G_i$ and $G^*$, $T_{G^*} - T_{G_i}$ and generate
              the acceptance ratio
                  $r = \exp\left(\theta^* \cdot (T_{G^*} - T_{G_i})\right);$
        else
            $p_2 > p_{edge}$: Toggle random dyad:
            Choose one of the possible dyads at random and change it from an edge to a non-edge if it is
              already in the network, and from a non-edge to an edge if it is not. This creates a proposed set
              of edges $Y^*$, and proposed network $G^* = (Y^*, X_i)$.
            We now have two networks, the current network $G_i$ and the proposed one $G^*$. We calculate the
              change statistics between $G_i$ and $G^*$, $T_{G^*} - T_{G_i}$, and generate the acceptance ratio
                  $r = \exp\left(\theta^* \cdot (T_{G^*} - T_{G_i})\right);$
        end
    else
        $p_1 > p_{dyad}$. We toggle element in $X$:
        Choose one node in the network at random, and then one of the attributes in $X$ to toggle at random.
          As mentioned at the beginning of this chapter, we are only interested in binary attributes.
          Toggling an binary attribute means changing it's value from 1 to 0, or vice versa. This creates a
          proposed set of attributes $X^*$, and proposed network $G^* = (Y_i, X^*)$.
        We now have two networks, the current network $G_i$ and the proposed one $G^*$. We calculate the
          change statistics between $G_i$ and $G^*$, $T_{G^*} - T_{G_i}$, and generate the acceptance ratio
              $r = \exp\left(\theta^* \cdot (T_{G^*} - T_{G_i})\right);$
    end
    Draw a random number from a uniform distribution $u$, $u \sim U(0,1)$;
    if $u < r$ then
        we assign the proposed network $G^*$ to $G_{i+1}$ for the next iteration;
    else
        we assign the current network $G_i$ to $G_{i+1}$ for the next iteration;
    end
end
```

1 inside of Algorithm 2 refers to a symmetric proposal $h(\theta_i)$, based on the value of θ at step i. Following Caimo and Friel (2011), $h(\cdot)$ is an arbitrary distribution to generate a proposal for θ_{i+1} based on the current value θ_i. In this, I use a random walk distribution centred at θ_i. In step 2, the algorithm creates a θ-ratio from the probability of the proposal for θ, θ^*, and the probability of current value of θ_i. This likelihood for which the simulated data are defined, and is sampled via an McMC using Algorithm 1. The practical implementation suggests using an weakly informative multivariate normal distribution.

```
Algorithm 2: Parameter estimation algorithm
Input: Observed network G=(Y, X), set of sufficient statistics to explain the network, $T_G$
Output: Distribution of parameters $\theta$ for observed network $G$
Consider a starting set for $\theta$, $\theta_0$. $\theta_0$ can be either all zeros or the value obtained by
  estimating the network using the maximum pseudolikelihood (as described in
  Section 2.2.1).;
Select a desired number of iterations, $l_2$, and burn-in $b$. ;
while $i < l_2$ do
    1. Draw $\theta^*$ from a symmetric proposal $h(\theta_i)$;
    2. Create $\theta$-ratio, defined as $pr = \frac{\text{prior}(\theta^*)}{\text{prior}(\theta_i)}$;
    3. Sample a network $G^* = (Y^*, X^*)$ using Algorithm 1, conditional on $\theta^*$ through $T_G(\cdot)$
      ;
    4. Calculate the change in statistics from the proposed network $T_G(G^*)$ and the
      observed network $T_G(G)$, as $\delta = T_G(G^*) - T_G(G)$;
    5. Using the above, determine the chance of accepting the new proposed $\theta^*$ as
      $\alpha = (\theta^* - \theta_i)\delta + \log(pr)$ where $\theta_i$ is the current value of $\theta$;
    6. Draw a random number from a uniform distribution $u \sim U(0, 1)$;
    if The logarithm of the random number $u$, $\log u$ is smaller than the proposal $\alpha$, $u < \alpha$
      then
        We assign the proposed value of $\theta^*$ to the next iteration $\theta_{i+1}$;
    else
        We assign the current value, $\theta_i$ to the next iteration $\theta_{i+1}$;
    end
    if The current iteration of the algorithm is larger than the burn-in $b$, $i > b$ then
        Save $\theta_{i+1}$ into a vector as part of the vector of parameters that will become samples
          from the posterior distribution;
    end
end
```

Notice how in Algorithm 2 the calculation of α is the log of the probability described in Equation 2.9, following Caimo and Friel (2014). The choice of the burn-in period is affected by the starting point of θ^* in the algorithm. Following the assumption that the pseudolikelihood estimation is close to the true value of θ, less burn-in is necessary to reach a stable posterior distribution.

5.3.3 Change statistics in the exponential random network model

In Subsection 2.2.1 I explained the concept of change statistics. In the context of an McMC routine, change statistics is a technique that reduces the amount of computing time needed to calculate the acceptance ratio between the vector of sufficient statistics of proposed network T_{G^*} and those of a current one T_{G_i}, $\exp\left(\theta^* \cdot (T_{G^*} - T_{G_i})\right)$. In the sampling methodology that generates networks by toggling dyads (whether existing or not), calculating the change statistics requires calculating how each individual statistic changes with the inclusion or exclusion of an edge. For many sufficient statistics, this is a relatively straightforward calculation. When calculating the change statistics for the sampling methodology that generates networks by toggling both edges *and* attributes, we need to go through a similar process that is specific to each one of the statistics (Li, 2015).

5.4 Estimation

To showcase the estimation strategy that Algorithm 2 produces, I am going to explore a commonly used ERGM example, the Sampson's "liking" dataset, which allows us to compare our estimation procedure with a classical directed network. Let us look initially at the results from using the traditional ergm assumptions used in the `statnet` estimation routine, as well as the one available in the `bergm` package (Caimo and Friel, 2014). The sampler described in Algorithm 1 can replicate the `statnet` sampler by not toggling node attributes but only toggling one dyad at a time. This is referred to as the "edge/dyad sampler"[1], because at every step of the McMC chain, the proposed network changes in only one connection. The "edge/dyad" name makes reference to the fact that the sampler will consider toggling existing edges with probability p_{edge} or dyads with a probability $1 - p_{edge}$ chance[2]. I use $1 - p_{dyad} = 20\%$ to encourage exploration of all the possible dyads, as

[1] This sampler is sometimes referred to as the tie/no tie sampler, but for consistency, I call it the dyad/edge sampler.

[2] See `https://rdrr.io/cran/ergm/man/ergm-proposals.html`

described in Subsection 5.3 . The dataset contains nominations of friendships between monks for three different moments in time (Sampson, 1969). A nomination made by monk *A* to monk *B* implies monk *A* likes monk *B*.

We are going to estimate the parameters for a model that considers the number of network connections (edges) and the number of reciprocated connections (mutual) as the sufficient statistics that characterise the network. The results generated by four estimation methods can be found in Table 5.1. The first one is the traditional `statnet` McMCMLE estimation routine, the second is the `bergm` estimation routine. The third one is the estimation using the custom "edge/dyad" sampler built for this project ("Custom E/D"), and the fourth one is uses "dyad-attribute" sampler ("Custom D/A"). The results for all four are relatively similar, considering that there are no node attributes considered in the set of sufficient statistics, so toggling node attributes should not change the probability of accepting a particular network proposal, and this is exactly what Table 5.1 shows.

Table 5.1: Estimation results for Sampson monk dataset for four different estimation methodologies for the following network model: *network* ~ *edges* + *mutual*

Estimation method	**Edges** Estimate (Standard Error)	**Mutual** Estimate (Standard Error)
`statnet`	-2.11 (0.21)	2.15 (0.47)
	Estimate (Standard Deviation)	Estimate (Standard Deviation)
`bergm`	-2.09 (0.22)	2.09 (0.51)
Custom E/D	-2.29 (0.21)	2.46 (0.38)
Custom D/A	-2.04 (0.17)	1.86 (0.36)

From Table 5.1, and basically all estimation procedures, we can see that there is a significant mutuality effect at play in the monk network. The fact that the edges term and the mutual term have similar but opposing values, means that the conditional log-odds of a new connection between two monks is 0, which translates into a probability of around 50%. Connections that are not mutual (represented by the edges term alone), indicate that the conditional log-odds of one monk to create a non-mutual connection with another monk is of around -2.11, or around 10%s.

We are now going to explore how these four estimation methodologies recover a set of parameters selected to produce non-degenerate networks. Table 5.1 shows that both custom estimation strategies developed for this dissertation produce results that are similar to the `statnet` and `bergm` estimations.

We can expand this example to include a node attribute belonging to the original dataset to compare the result from the four estimation described strategies. The data collected by Sampson includes whether the monks attended the minor seminary "Cloisterville" before joining the monastery where the data was collected. The statistic we are adding to the estimation is often referred to as `nodematch`, and it measures the extent to which two individuals who are connected in the network share the same attribute value. The estimates are presented in Table 5.2. The results show that including a statistic that considers the connections between the monks that probably knew each other from before decreases the level of mutuality in the network.

Table 5.2: Estimation results for Sampson monk dataset for four different estimation methodologies for the following network model: $network \sim edges + mutual + nodematch(Cloisterville)$

Estimation method	**Edges** Estimate (Standard Error)	**Mutual** Estimate (Standard Error)	**Nodematch("Cloisterville")** Estimate (Standard Error)
`statnet`	-1.96 (0.23)	2.08 (0.47)	-0.27 (0.24)
	Estimate (Standard Deviation)	Estimate (Standard Deviation)	Estimate (Standard Deviation)
`bergm`	-1.98 (0.26)	2.11 (0.51)	-0.25 (0.26)
Custom E/D	-1.97 (0.24)	1.92 (0.43)	-0.23 (0.25)
Custom D/A	-2.04 (0.25)	2.03 (0.50)	-0.18 (0.24)

Table 5.1 and 5.2 show that the estimation procedure developed for this chapter produces results comparable to the existing software implementations. I am now going to show using simulations that this procedure can recover a vector of parameters θ^* that is used to generate random networks $Y = y, X = x|\theta^*$.

5.4.1 Simulation

In Section 5.3 I showed how the sampling algorithm works to produce network samples conditional on a parameter of θ's, depending on the desired data generating process. Section 5.4 showed that the coupling of the network sampling methodology with the estimation procedure based on Caimo and Friel (2011) produces results comparable to the ERGM estimation procedure. I will now show that the estimation procedure developed for this chapter recovers the true values for networks generated using the custom sampler.

Let us start with a known value of θ^* which we will consider the truth and will try to recover. This vector needs to generate admissible networks - networks that we are likely to observe in real life, or those that at least, are not degenerate. For this we are going to consider a visual inspection enough. The simulation procedure works as follows. For a fixed set of simulations:

1. Given the vector of true θ^*, sample a network using the "dyad and attribute toggle".

2. Estimate the observed network using both the ERGM and BERGM estimation routines developed as a benchmark for the standard dyad-only toggle sampler. Additionally, estimate the observed network using the "dyad and attribute" toggle sampler, and save the coefficients. In the case of the ERGM, save the point estimate and standard errors produced by the estimation. In the case of the BERGM and the custom sampler, save the mean of the distribution, as well as the standard deviation.

3. Aggregate the estimation results for all simulations, and calculate the difference to the vector of true parameters, θ^*.

We are going to generate networks, with nodes that have one binary attribute, according to the following network model:

$$G = (Y, X) \text{ edges} + \textit{mutual} + \text{nodematch},$$

and the following vector of covariates: $\theta^* = \{-1, -0.2, 0.5\}$. This parameter vector was considered because of how the conditional log-odds translate into network features. One sample network generated using this set of sufficient statistics looks like the network in Figure 5-1. The results from the simulation are presented in Table 5.3. All three estimation procedures come close to recovering the true values of θ^*. This simulation can be extended to explore a larger area of the parameter space, and determine whether there are networks generated with a specific vector of parameters where the estimates differ. This is, that the estimation using a sampler that toggles dyads and attributes is closer to the true values than a sampler that toggles both dyads and attributes.

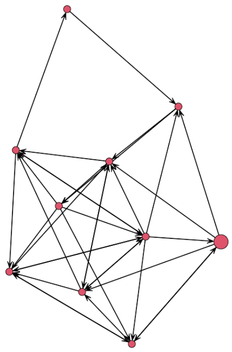


Figure 5-1: Example of sample generated network

Table 5.3: Average estimate results for 100 simulations using three different estimation methodologies using the following network model: *network* ∼ *edges* + *mutual* + *nodematch*

Statistic	θ^*	`statnet` Estimate (Standard Error)	`bergm` Estimate (Standard Deviation)	Custom D/A Estimate (Standard Deviation)
Edges	-1	-0.99 (0.50)	-1.03 (0.51)	-0.98 (0.42)
Mutual	-0.2	-0.35 (0.73)	-0.44 (0.51)	-0.38 (0.74)
Nodematch	0.5	0.56 (0.57)	0.62 (0.58)	0.40 (0.30)

5.5 Conclusion

In this chapter I presented a parameter estimation methodology that relies on a more general network sampling procedure than the one considered by traditional parameter estimation routines for the exponential random graph model. This sampler originally appeared in Fellows and Handcock (2012), who used it with a maximum likelihood estimation methodology to characterise observed networks. The possibility of toggling node attributes increases the range of cases to be considered for analysis, since it incorporates a large amount of networks in the calculation of the normalising constant. In relation to the Bayesian approach, it means that the sampler navigates a portion of the parameter space that generates viable (non-degenerate) candidates in each step of the McMC chain. Assuming the node attributes of an observed network as static and exogenous, reduces the space of networks to be explored in estimation.

The procedure follows the blueprint from the estimation strategy developed by Caimo and Friel (2011): a Bayesian framework that uses Markov chain Monte Carlo (McMC) algorithms to estimate the vector of parameters corresponding to the vector of sufficient statistics selected to explain the observed network. Caimo and Friel mention that their methodology (back in 2011) was an initial step in understanding the interplay between Bayesian estimation and exponential random graph models. This chapter aimed to increase the number of networks that can be analysed with their approach.

From Section 5.3.2 we can see that the proposed estimation methodology generates results that are equivalent to those by the maximum likelihood estimation from the `statnet` suite (Krivitsky et al., 2020), and the `bergm` package (Caimo and Friel, 2014). This means that, for the scenario explored, the dyad/attribute sampler replicates the vector of parameters when including and excluding network related variables. In addition to that, Section 5.4.1 showed that the estimation procedure also recovers true parameters in a simulation setting.

This is a first step in a more general Bayesian estimation routine that will allow researchers to determine whether attributes coevolved with the formation of networks when the network and the attributes were observed through time. In the example presented in the introduction to this chapter, it would be possible to determine whether the uptake of smoking by some individuals happened after they made connections with individuals that also smoked. In addition to this future development, more research is needed to understand where the estimation produced by these two samplers diverges. This is, when considering networks whose data generating process involves attributes that might change in the process of network formation, potentially resulting in biased estimates from failing to consider the correct data-generating process behind the networks that are being analysed.

Chapter 6

Discussion and conclusions

I set out this dissertation with the aim of answering the question: "how to determine the causal relationship between different variables with interdependent observations?". As explained in the Introduction (Chapter 1), the interdependence mentioned in that question makes reference to the fact that the considered units of observation are not independent from each other as it is usually assumed with more traditional statistical methodologies. I explored this question in three different chapters which consider three different kinds of interdependence between the observations.

In Chapter 3, I presented a methodology that aims to estimate the causal direct effect of an intervention, as well as the causal spillover effect of the intervention in the case where the treated units exposed the control units that were not originally assigned to treatment. Traditional causal inference methodologies look to estimate the causal relationship (and not just the correlation) between two variables, usually a treatment and an outcome variable, properly discounting the effect other variables might have on the outcome. In this case we are looking at a treatment applied to a group of individuals selected to be treated and an observed outcome by those individuals. Assuming that some of the treated units expose those not originally intended to receive treatment further complicates the challenge of getting unbiased causal estimates for the effect of the intervention since it is now crucial to consider the exposure to treatment from neighbouring units.

Forastiere et al. (2020) proposed a methodology that allows researchers to separate the effect of the intervention into causal direct and spillover effects, conditional on a set of strict but plausible assumptions. Their method uses generalised propensity scores to calculate the average potential outcomes of being assigned to a particular kind of treatment, and being exposed to a different one. I explore this methodology in detail and find that without perfect information about the outcome model, their estimation routine produces biased results. As an alternative, I propose using a flexible regression based on cubic splines that does not require full knowledge of the outcome model and produces relatively less biased results. In addition, I use the methodology developed by Forastiere et al. (2020) as well as our own methodology to estimate the causal direct and spillover effects of the Sexually Transmitted infections And Sexual Health (STASH) intervention and find that although there was a positive impact on the treated individuals, the outcome of the individuals exposed to the intervention did not behave in the same way.

In Chapter 4 I explored the limits of randomisation inference as a tool to causally determine the effect order had in the sequence of events organised as a two-mode network. In particular for this dissertation, I focus on exploring two-mode networks where the nodes in the first mode sequentially interact with the nodes in the second mode, and refer to them as bipartite event cascades. To characterise the nature of this sequence of connections I use the relational event model framework developed by Butts (2008), and with it, estimate a homophily statistic that captures to what extent senders (nodes in the first mode) behave similarly to other senders that have interacted with the same receivers (actors in the second mode) in the past.

Randomisation inference is used to determine how important the order is in determining the way senders connect to receivers. Rejection of the null hypothesis that order does not matter indicates that the succession of events was important (not at random). Careful consideration of the different causal pathways that might lead to this outcome suggests the presence of influence between the senders of the two-mode network. The idea of using randomisation inference for this purpose has been explored in a practical application before (see Malang et al. 2019), however, it is unclear under what conditions the methodology correctly estimates the extreme nature of the original coefficient. To corroborate our theoretical results, I use a set of simulations that produce event sequences compatible with bipartite event cascades. The results suggest that there is an optimal area in the space of parameters (number of senders, number of receivers, number of events in the sequence, among others) where randomisation inference correctly determines the presence of influence in the event sequence, and that deviations from that area suggest more or less influence than is actually present in the sequence.

In Chapter 5 I introduced a new estimation procedure for the exponential random network model (ERNM) originally developed by Fellows and Handcock (2012). Traditional exponential random graph models (ERGM) estimate a set of parameters that, with a set of selected sufficient statistics, represent the likelihood of observing a particular configuration of nodes and their connections. The most traditional estimation methodology uses maximum likelihood estimation, where the normalising function is approximated from a collection of sampled networks. These samples are generated by toggling the connections between the nodes, but assuming that the node attributes remain static. Caimo and Friel (2011) developed an estimating methodology that uses Bayesian analysis to produces a set of parameters that explain the observed network, however, it relies on the same sampler that only toggles edges between the nodes to produce new candidate networks.

Fellows and Handcock (2012) introduced a new sampling methodology where both the edges and the node attributes are toggled to produce new networks, and their estimation methodology was based on maximum likelihood estimation. I constructed a sampling methodology that incorporates the node covariates as endogenous parts of the model and toggles them as well as the network edges. Instead of using MLE, which as suggested by the authors, is a methodology prone to producing degenerate networks, I implemented the Bayesian approach developed by Caimo and Friel (2011), in order to produce estimates with a more nuanced degree of uncertainty, but also because the algorithm tends to avail more the areas of the parameter space that produce degenerate results. The resulting estimation procedure matches the results produced by the `statnet` and `bergm` methodologies, but increases the range of networks considered to generate the estimation. This means that it could potentially be used to flesh out biases in estimation when the data generating process underlying a network has endogenous nodal attributes.

In all three chapters I aimed to present a contribution to existing methodologies that consider units of observation with a certain degree of interdependence. In Chapter 3 I expanded on a methodology that estimates the causal effect of an intervention in the presence of treatment interference. In Chapter 4 I explored the limits of a tool used to determine the extent of interdependence in bipartite event cascades. In Chapter 5 I introduced a new estimation methodology that considers a more general data generating process by sampling networks where both edges and node attributes vary, and uses recent developments in Bayesian analysis to reduce the number of degenerate samples considered for the estimation.

Interdependent data in statistical research poses a challenge to the probabilistic underpinning of the field in that it violates the commonly used assumption that units of observation are random draws independent from each other. One of the implications of relaxing this hypothesis is that it requires the analyst to have, at least, an idea of how the units of observation are connected to each other. In the case of bipartite event cascades, it means knowing the order in which the cascade occurred, to then pry it apart to test whether the originally observed statistic is extreme or not. When using randomisation inference and bipartite event sequences, I assume that the event sequences are not directly modified by the starting attributes of the senders or the receivers (see Figure 4-2). This assumption could be relaxed and an additional step could be added to the simulations to reflect this. There are currently no theoretical explanations as to what would happen, other than the fact that such a causal mechanism probably violates the no-confoundedness assumptions necessary for causal identification. Future research should focus on understanding how causally confounded estimates behave when analysed using the shuffle test.

For public health interventions, understanding this interdependence means that we can consider trial designs different than the more traditional ones where everyone selected to be treated, actually gets treated, and reduce bias estimating the effect of the intervention as shown in Chapter 3. One of the first questions that come up in relation to future work asks about the equivalence in terms of average treatment effect between treating all units in a sample and treating some of these units and expecting the treatment to spill over to the units that were not treated. If that equivalence exists, to what extent does it depend on the kinds of connections between the individuals (friendship, mentorship, admiration), and the way in which those connections were formed (at random, with a strong homophily pattern, considering triadic closure).

One of the motivating questions of this research, mentioned in Chapter 1, is whether selection (the fact that individuals create connections with others similar to them) can be disentangled from influence (the fact that individuals adapt their behaviour based on their connections) in social network studies. Without proper information and strong assumptions, this is, actually, impossible (Shalizi and Thomas, 2011). In the context of causal inference in the presence of interference, this question is relevant because we assume that the connections between the students are a given and static. Aside from the estimation of the neighbourhood propensity score that determines how likely it is that units get exposed to treatment, there are no additional considerations on the actual structure of the network. This is by design, since the logistic regressions used to generate this propensity score cannot account for complex network interdependencies in the same way the exponential random graph/network models do.

In contrast, the exponential random network model developed by Fellows and Handcock (2012), and by extension, the one developed for this dissertation, allows for these kinds of complex interdependencies. ERNMs allow the researcher to explore the probability of observing a specific configuration of connections and a set of attributes, $P(Y = y, X = x|\theta)$. Fellows and Handcock show evidence of how there is bias in estimating the correlation between a node attribute and a specific outcome when entirely ignoring network structure. Being able to include the structural information on the connections between individuals, in conjunction with ideas behind causal inference, is the natural methodological extension to the developments presented in this dissertation.

In addition to this, being able to generate networks to understand a specific matrix of node attributes given the way they are connected to each other, $P(X = x|\theta, Y = y)$, means that the ERNM model can be considered as an alternative estimating strategy to the network autocorrelated model (NAM) (see Dittrich et al. (2017) for more on Bayesian estimation of temporal NAMs).

Bibliography

Anagnostopoulos, Aris, Ravi Kumar, and Mohammad Mahdian (2008). Influence and correlation in social networks. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 7–15.

Angrist, Joshua D. and Jörn-Steffen Pischke (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

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

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

Athey, Susan, Dean Eckles, and Guido W. Imbens (2018). Exact p-values for network interference. Journal of the American Statistical Association 113(521), 230–240.

Athey, Susan and Guido W. Imbens (2017). The state of applied econometrics: causality and policy evaluation. The Journal of Economic Perspectives 31(2), 3–32.

Barankin, Edward W. and Ashok P. Maitra (1963). Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics. Sankhya: The Indian Journal of Statistics, Series A (1961-2002) 25(3), 217–244.

Barndorff-Nielsen, Ole E. (1978). Information and exponential families: In statistical theory. John Wiley & Sons, Ltd.

Bernauer, Thomas, Anna Kalbhenn, Vally Koubi, and Gabriele Spilker (2010). A comparison of international and domestic sources of global governance dynamics. British Journal of Political Science 40(3), 509–538.

Blossfeld, Hans-Peter and Götz Rohwer (2001). Techniques of event history modeling: new approaches to casual analysis (2nd edition ed.). Psychology Press.

Brandenberger, Laurence (2019). Predicting network events to assess goodness of fit of relational event models. Political Analysis 27(4), 556–571.

Brandes, Ulrik, Jürgen Lerner, and Tom A. B. Snijders (2009). Networks evolving step by step: Statistical analysis of dyadic event data. In 2009 International Conference on Advances in Social Network Analysis and Mining, pp. 200–205. IEEE.

Brown, Lawrence D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. Lecture Notes-Monograph Series 9, i–279.

Butts, Carter T. (2001). The complexity of social networks: Theoretical and empirical findings. Social Networks 23(1), 31–72.

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

Butts, Carter T. (2011). Bernoulli graph bounds for general random graphs. Sociological Methodology 41, 299–345.

Butts, Carter T., David Hunter, Mark Handcock, Skye Bender-deMoll, Jeffrey Horner, Li Wang, Pavel N. Krivitsky, Brendan Knapp, Michał Bojanowski, and Chad Klumb (2021). network: Classes for Relational Data.

Caimo, Alberto and Nial Friel (2011). Bayesian inference for exponential random graph models. Social Networks 33(1), 41–55.

Caimo, Alberto and Nial Friel (2014). Bergm: Bayesian Exponential Random Graphs in R. Journal of Statistical Software 61, 1–25.

Campbell, Benjamin W., Frank W. Marrs, Tobias Böhmelt, Bailey K. Fosdick, and Skyler J. Cranmer (2019). Latent influence networks in global environmental politics. PloS ONE 14(3).

Campbell, Rona, Fenella Starkey, Joe Holliday, Suzanne Audrey, Michael Bloor, Nina Parry-Langdon, Rachael Hughes, and Laurence Moore (2008). An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. Lancet 371(9624), 1595–1602.

Carrington, Peter J., John Scott, and Stanley Wasserman (2005). Models and Methods in Social Network Analysis. Cambridge, U.K.; New York: Cambridge University Press.

Chen, C., I. Noble, J. Hellmann, J. Coffee, M. Murrillo, and N. Chawla (2015). University of Notre Dame global adaptation index. Country index technical report, University of Notre Dame. `https://gain.nd.edu/assets/254377/nd_gain_technical_document_2015.pdf`. Accessed 28 February 2021.

Christakis, Nicholas A. and James H. Fowler (2007). The spread of obesity in a large social network over 32 years. New England Journal of Medicine 357(4), 370–379.

Cox, D. R. (1958). Planning of experiments. New York: Wiley.

Cranmer, Skyler J., Bruce A. Desmarais, and Elizabeth J. Menninga (2012). Complex dependencies in the alliance network. Conflict Management and Peace Science 29(3), 279–313.

Cranmer, Skyler J., Philip Leifeld, Scott D. McClurg, and Meredith Rolfe (2017a). Navigating the range of statistical tools for inferential network analysis. American Journal of Political Science 61(1), 237–251.

Cranmer, Skyler J., Philip Leifeld, Scott D. McClurg, and Meredith Rolfe (2017b). Navigating the range of statistical tools for inferential network analysis. American Journal of Political Science 61(1), 237–251.

Cunningham, Scott (2021). Causal Inference: The Mixtape. New Haven ; London: Yale University Press.

Daum, F. E. (1986). The Fisher-Darmois-Koopman-Pitman theorem for random processes. In 1986 25th IEEE Conference on Decision and Control, pp. 1043–1044.

Dekker, David, David Krackhardt, and Tom A. B. Snijders (2007). Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. Psychometrika 72(4), 563–581.

Del Prete, Davide, Laura Forastiere, and Valerio Leone Sciabolazza (2020). Causal inference on networks under continuous treatment interference. arXiv:2004.13459 [econ, stat]. arXiv: 2004.13459.

Desmarais, B. A. and S. J. Cranmer (2012). Statistical mechanics of networks: Estimation and uncertainty. Physica A: Statistical Mechanics and its Applications 391(4), 1865–1876.

Desmarais, Bruce A., Jeffrey J. Harden, and Frederick J. Boehmke (2015). Persistent policy pathways: Inferring diffusion networks in the American states. American Political Science Review 109(2), 392–406.

Dittrich, Dino, Roger T. A. J. Leenders, and Joris Mulder (2017). Bayesian estimation of the network autocorrelation model. Social Networks 48, 213–236.

Efron, Bradley (1977). The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association 72(359), 557–565.

Efron, Bradley (1978). The geometry of exponential families. The Annals of Statistics 6(2), 362–376. Publisher: Institute of Mathematical Statistics.

Elkins, Zachary, Andrew T. Guzman, and Beth A. Simmons (2006). Competing for capital: The diffusion of bilateral investment treaties, 1960–2000. International Organization (4), 811–846.

Engber, Daniel (2012). The Internet Blowhard's Favorite Phrase. Slate. Section: Science.

Entner, Doris, Patrik Hoyer, and Peter Spirtes (2013). Data-driven covariate selection for nonparametric estimation of causal effects. In Carlos M. Carvalho and Pradeep Ravikumar (Eds.), Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, Volume 31 of Proceedings of Machine Learning Research, Scottsdale, Arizona, USA, pp. 256–264. PMLR.

Erdős, P. and A Rényi (1960). On the evolution of random graphs. In Publication of the Mathematical Institute of the Hungarian Academy of Sciences, pp. 17–61.

Fellows, Ian and Mark S. Handcock (2012). Exponential-family random network models. arXiv:1208.0121 [stat]. arXiv: 1208.0121.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transantions Royal Society of London. Series A 222(594-604), 309–368.

Fisher, Ronald Aylmer (1935). The design of experiments. New York: Hafner Press.

Flache, Andreas, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz (2017). Models of social influence: towards the next frontiers. Journal of Artificial Societies and Social Simulation 20(4), 2.

Forastiere, Laura, Edoardo M. Airoldi, and Fabrizia Mealli (2020). Identification and estimation of treatment and interference effects in observational studies on networks. Journal of the American Statistical Association 116(534), 901–918.

Forsyth, Ross, Carrie Purcell, Sarah Barry, Sharon Simpson, Rachael Hunter, Lisa McDaid, Lawrie Elliot, Julia Bailey, Kirsty Wetherall, Mark McCann, Chiara Broccatelli, Laurence Moore, and Kirstin Mitchell (2018). Peer-led intervention to prevent and reduce STI transmission and improve sexual health in secondary schools (STASH): protocol for a feasibility study. Pilot and Feasibility Studies 4(1), 180.

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

Fowler, James H. (2006). Connecting the Congress: A study of cosponsorship networks. Political Analysis 14(4), 456–487.

Fowler, James H. and Nicholas A. Christakis (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. BMJ 337, a2338.

Frank, Ove and David Strauss (1986). Markov graphs. Journal of the American Statistical Association 81(395), 832–842.

Franzese, Jr., Robert J. and Jude C. Hays (2007). Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. Political Analysis 15(2), 140–164.

Freedom House (2020). Freedom in the world. `https://freedomhouse.org/report/freedom-world`. Accessed 28 February 2021.

Geyer, Charles J. (1991). Markov Chain Monte Carlo Maximum Likelihood. Interface Foundation of North America. Accepted: 2010-02-24T20:38:06Z.

Geyer, Charles J. (2021). Exponential families. Department of Statistics, University of Minnesota. Lecture.

Geyer, Charles J. and Elizabeth A. Thompson (1992). Constrained Monte Carlo maximum likelihood for dependent data. Journal of the Royal Statistical Society. Series B (Methodological) 54(3), 657–699.

Gilbert, E. N. (1959). Random graphs. The Annals of Mathematical Statistics 30(4), 1141–1144.

Gleditsch, Kristian Skrede and Michael D. Ward (2006). Diffusion and the international context of democratization. International Organization 60(4), 911–933.

Gomez-Rodriguez, Manuel, Jure Leskovec, and Andreas Krause (2012). Inferring networks of diffusion and influence. ACM Transactions on Knowledge Discovery from Data (TKDD) 5(4), 1–37.

Gomez-Rodriguez, Manuel, Jure Leskovec, and Bernhard Schölkopf (2013). Modeling information propagation with survival theory. In International Conference on Machine Learning, pp. 666–674.

Hall, Peter, Jeff Racine, and Qi Li (2004). Cross-validation and the estimation of conditional probability densities. Journal of the American Statistical Association 99(468), 1015–1026.

Halloran, M. Elizabeth and Michael G. Hudgens (2012). Causal inference for vaccine effects on infectiousness. The International Journal of Biostatistics 8(2), 1–40.

Halloran, M. Elizabeth and Michael G. Hudgens (2016). Dependent happenings: a recent methodological review. Current Epidemiology Reports 3(4), 297–305.

Handcock, Mark S (2003). Assessing degeneracy in statistical models of social networks. pp. 27. Working Paper no. 39. Center for Statistics and the Social Sciences. University of Washington.

Hanneke, Steve, Wenjie Fu, and Eric P. Xing (2010). Discrete temporal models of social networks. Electronic Journal of Statistics 4, 585–605.

Harden, Angela, Ann Oakley, and Sandy Oliver (2001). Peer-delivered health promotion for young people: A systematic review of different study designs. Health Education Journal 60(4), 339–353.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1), 97–109.

Hernán, Miguel A and James M Robins (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

Hirano, Keisuke and Guido W. Imbens (2005). The propensity score with continuous treatments. In Andrew Gelman and Xiao-Li Meng (Eds.), Wiley Series in Probability and Statistics, pp. 73–84. Chichester, UK: John Wiley & Sons, Ltd.

Hirvonen, Maija, Carrie Purcell, Lawrie Elliott, Julia V. Bailey, Sharon Anne Simpson, Lisa McDaid, Laurence Moore, Kirstin Rebecca Mitchell, and The STASH Study Team (2021). Peer-to-peer sharing of social media messages on sexual health in a school-based intervention: opportunities and challenges identified in the STASH feasibility trial. Journal of Medical Internet Research 23(2), e20898.

Ho, Kate and Adam M Rosen (2015). Partial identification in applied research: Benefits and challenges. Working Paper 21641, National Bureau of Economic Research.

Holland, Paul W. and Samuel Leinhardt (1981). An Exponential Family of Probability Distributions for Directed Graphs. Journal of the American Statistical Association 76(373), 33–50. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

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

Hugh-Jones, David, Karolina Milewicz, and Hugh Ward (2018). Signaling by signature: The weight of international opinion and ratification of treaties by domestic veto players. Political Science Research and Methods 6(1), 15–31.

Hunter, David R and Mark S Handcock (2006). Inference in curved exponential family models for networks. Journal of Computational and Graphical Statistics 15(3), 565–583.

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks.

Hunter, David R., Pavel N. Krivitsky, and Michael Schweinberger (2012). Computational statistical methods for social network models. Journal of Computational and Graphical Statistics 21(4), 856–882.

Jackson, Matthew O. (2010). Social and Economic Networks. Princeton, NJ: Princeton University Press.

Jagadeesan, Ravi, Natesh S. Pillai, and Alexander Volfovsky (2020). Designs for estimating the treatment effect in networks with interference. The Annals of Statistics 48(2).

Janzing, Dominik (2007). On causally asymmetric versions of Occam's Razor and their relation to thermodynamics. arXiv:0708.3411 [cond-mat, physics:quant-ph]. arXiv: 0708.3411.

Kang, Hyunseung and Guido Imbens (2016). Peer encouragement designs in causal inference with partial interference and identification of local average network effects. arXiv:1609.04464 [stat]. arXiv: 1609.04464.

Keele, Luke (2015). The statistics of causal inference: A view from political methodology. Political Analysis 23(3), 313–335.

Kirby, Douglas B., B. A. Laris, and Lori A. Rolleri (2007). Sex and HIV education programs: their impact on sexual behaviors of young people throughout the world. The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine 40(3), 206–217.

Kleinberg, Jon (2007). Cascading behavior in networks: Algorithmic and economic issues. Algorithmic Game Theory 24, 613–632.

Kolaczyk, Eric D. (2009). Statistical analysis of network data. Springer Series in Statistics. Springer New York. DOI: 10.1007/978-0-387-88146-1.

Kolaczyk, Eric D. (2017). Topics at the Frontier of Statistics and Network Analysis: (Re)Visiting the Foundations. ISBN: 9781108290159 9781108407120 Publisher: Cambridge University Press.

Koskinen, Johan (2004). Bayesian analysis of exponential random graphs - estimation of parameters and model selection. Research Report, University of Manchester.

Koskinen, J. (2008). The linked importance sampler auxiliary variable metropolis hastings algorithm for distributions with intractable normalising constants. MelNet Social Networks Laboratory Technical Report.

Koskinen, Johan H., Garry L. Robins, and Philippa E. Pattison (2010). Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. Statistical Methodology 7(3), 366–384.

Krackardt, David (1987). QAP partialling as a test of spuriousness. Social Networks 9(2), 171–186.

Krivitsky, Pavel N., Mark S. Handcock, David R. Hunter, Carter T. Butts, Chad Klumb, Steven M. Goodreau, and Martina Morris (2003-2020). statnet: Software tools for the Statistical Modeling of Network Data. Statnet Development Team.

Krivitsky, Pavel N., Mark S. Handcock, Adrian E. Raftery, and Peter D. Hoff (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. Social Networks 31(3), 204–213.

La Fond, Timothy and Jennifer Neville (2010). Randomization tests for distinguishing social influence and homophily effects. In Proceedings of the 19th International Conference on World Wide Web, pp. 601–610.

Lee, Youjin and Elizabeth L. Ogburn (2020). Network dependence can lead to spurious associations and invalid inference. Journal of the American Statistical Association 116(535), 1060–1074.

Lee Ray, James (2013). War on democratic peace. International Studies Quarterly 57(1), 198–200.

Leenders, Roger Th A. J. (1997). LONGITUDINAL BEHAVIOR OF NETWORK STRUCTURE AND ACTOR ATTRIBUTES: MODELING INTERDEPENDENCE OF CONTAGION AND SELECTION. In Evolution of Social Networks. Routledge. Num Pages: 20.

Leifeld, Philip (2017). Discourse network analysis: Policy debates as dynamic networks. In Jennifer N. Victor, Alexander H. Montgomery, and Mark N. Lubell (Eds.), The Oxford handbook of political networks, pp. 301–325. Oxford: Oxford University Press.

Lerner, Jürgen, Margit Bussmann, Tom A. B. Snijders, and Ulrik Brandes (2013). Modeling frequency and type of interaction in event networks. Corvinus Journal of Sociology and Social Policy 4(1), 3–32.

Lerner, J¸rgen, Natalie Indlekofer, Bobo Nick, and Ulrik Brandes (2013). Conditional independence in dynamic networks. Journal of Mathematical Psychology 57(6), 275–283.

Lerner, Jürgen and Alessandro Lomi (2020). Reliability of relational event model estimates under sampling: How to fit a relational event model to 360 million dyadic events. Network Science 8(1), 97–135.

Li, Ke (2015). Degeneracy, Duration, and Co-evolution: Extending Exponential Random Graph Models (ERGM) for Social Network Analysis. Thesis.

Lusher, D., J. Koskinen, and G. Robins (2012). Exponential random graph models for social networks: theory, methods, and applications. Cambridge University Press.

Lyons, Russell (2011). The spread of evidence-poor medicine via flawed social-network analysis. Statistics, Politics and Policy 2(1), 1–26.

Malang, Thomas, Laurence Brandenberger, and Philip Leifeld (2019). Networks and social influence in European legislative politics. British Journal of Political Science 49(4), 1475–1498.

Malang, Thomas and Philip Leifeld (2021). The Latent Diffusion Network among National Parliaments in the Early Warning System of the European Union. JCMS: Journal of Common Market Studies 59(4), 873–890.

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

Marrs, Frank W., Benjamin W. Campbell, Bailey K. Fosdick, Skyler J. Cranmer, and Tobias Böhmelt (2019). Inferring influence networks from longitudinal bipartite relational data. Journal of Computational and Graphical Statistics, 1–13.

Meyn, Sean, Richard L. Tweedie, and Peter W. Glynn (2009). Markov chains and stochastic stability (2 ed.). Cambridge Mathematical Library. Cambridge University Press.

Morris, Martina, Mark S. Handcock, and David R. Hunter (2008). Specification of exponential-family random graph models: terms and computational aspects. Journal of Statistical Software 24(i04).

Murray, Iain, Zoubin Ghahramani, and David MacKay (2012). MCMC for doubly-intractable distributions. arXiv:1206.6848 [stat]. arXiv: 1206.6848.

Nielsen, Frank and Vincent Garcia (2009). Statistical exponential families: A digest with flash cards. arXiv:0911.4863 [cs]. arXiv: 0911.4863.

Ogburn, Elizabeth L. (2017). Challenges to estimating contagion effects from observational data. arXiv:1706.08440 [stat]. arXiv: 1706.08440.

Ogburn, Elizabeth L., Oleg Sofrygin, Ivan Diaz, and Mark J. van der Laan (2020). Causal inference for social network data. arXiv:1705.08527 [math, stat]. arXiv: 1705.08527.

Ogburn, Elizabeth L. and Tyler J. VanderWeele (2014). Causal diagrams for interference. Statistical Science 29(4), 559–578. arXiv: 1403.1239.

Paluck, Elizabeth Levy, Hana Shepherd, and Peter M. Aronow (2016). Changing climates of conflict: a social network experiment in 56 schools. Proceedings of the National Academy of Sciences 113(3), 566–571.

Papadogeorgou, Georgia, Fabrizia Mealli, and Corwin M. Zigler (2019). Causal inference with interfering units for cluster and population level treatment allocation programs. Biometrics 75(3), 778–787.

Pearl, Judea (1995). Causal diagrams for empirical research. Biometrika 82(4), 669–688.

Pearl, Judea (2009). Causality (2Rev e. edition ed.). Cambridge, U.K. ; New York: Cambridge University Press.

Perez-Heydrich, Carolina, Michael G. Hudgens, M. Elizabeth Halloran, John D. Clemens, Mohammad Ali, and Michael E. Emch (2014). Assessing effects of cholera vaccination in the presence of interference. Biometrics 70(3), 731–744.

Perrin, Sophie and Thomas Bernauer (2010). International regime formation revisited: Explaining ratification behaviour with respect to long-range transboundary air pollution agreements in Europe. European Union Politics 11(3), 405–426.

Rinaldo, Alessandro, Stephen E. Fienberg, and Yi Zhou (2009). On the geometry of discrete exponential families with application to exponential random graph models. Electronic Journal of Statistics 3, 446–484.

Ringe, Nils, Jennifer Nicoll Victor, and Justin H. Gross (2013). Keeping your friends close and your enemies closer? Information networks in legislative politics. British Journal of Political Science 43(3), 601–628.

Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher (2007). An introduction to exponential random graph (p*) models for social networks. Social Networks 29(2), 173–191.

Robins, Garry, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison (2007). Recent developments in exponential random graph (p*) models for social networks. Special Section: Advances in Exponential Random Graph (p*) Models 29(2), 192–215.

Rodriguez, Manuel Gomez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf (2014). Uncovering the structure and temporal dynamics of information propagation. Network Science 2(1), 26–65.

Rogers, Everett M. (2002). Diffusion of preventive innovations. Addictive Behaviors 27(6), 989–993.

Rosenbaum, Paul R. (2010). Design of observational studies. Springer Series in Statistics. New York: Springer-Verlag.

Rosenbaum, Paul R. and Donald B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70(1), 41–55.

Rougier, Jonathan (2017). Statistical Inference. University of Bristol. Lecture.

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

Rubin, Donald B. (1977). Assignment to Treatment Group on the Basis of a Covariate. Journal of Educational Statistics 2(1), 1–26.

Rubin, Donald B. (1980). Randomization analysis of experimental data: the Fisher randomization test (comment). Journal of the American Statistical Association 75(371), 591–593.

Rubin, Donald B. (1986). Comment: Which Ifs Have Causal Answers. Journal of the American Statistical Association 81(396), 961–962.

Rubin, Donald B. (2005). Basic concepts of statistical inference for causal effects in experiments and observational studies. Department of Statistics, Harvard University. Lecture.

Sampson, Samuel F. (1969). Crisis in a cloister. Ph. D. thesis, Cornell University.

Saul, Bradley C., Michael G. Hudgens, and M. Elizabeth Halloran (2017). Chapter 9 - Causal inference in the study of infectious disease. In Arni S. R. Srinivasa Rao, Saumyadipta Pyne, and C. R. Rao (Eds.), Handbook of Statistics, Volume 36 of Disease Modelling and Public Health, Part A, pp. 229–246. Elsevier.

Schweinberger, Michael, Pavel N. Krivitsky, Carter T. Butts, and Jonathan R. Stewart (2020). Exponential-family models of random graphs: Inference in finite, super and infinite population scenarios. Statistical Science 35(4), 627–662.

Schweinberger, Michael and Pamela Luna (2017). Hierarchical exponential-family random graph models. Journal of Statistical Software, Number of pages: 39.

Shalizi, Cosma Rohilla (2021). Advanced data analysis from an elementary point of view (unpublished manuscript). Carnegie Mellon University.

Shalizi, Cosma Rohilla and Alessandro Rinaldo (2013). Consistency under sampling of exponential random graph models. The Annals of Statistics 41(2), 508–535.

Shalizi, Cosma Rohilla and Andrew C. Thomas (2011). Homophily and contagion are generically confounded in observational social network Studies. Sociological methods & research 40(2), 211–239.

Shim, Jae-Mahn and Eunjung Shin (2020). Drivers of ratification rates in global biodiversity governance: Local environmentalism, orientation toward global governance, and peer pressure. Environmental Politics 29(5), 845–865.

Shipan, Charles R. and Craig Volden (2008). The mechanisms of policy diffusion. American Journal of Political Science 52(4), 840–857.

Simmons, Beth A. (2000). International law and state behavior: Commitment and compliance in international monetary affairs. American Political Science Review 94(4), 819–835.

Snijders, Tom, Christian Steglich, and Michael Schweinberger (2006). Modeling the coevolution of networks and behavior. In Longitudinal Models in the Behavioral and Related Sciences, pp. 31. Routledge.

Snijders, Tom A. B. (2001). The statistical evaluation of social network dynamics. Sociological Methodology 31(1), 361–395.

Snijders, Tom A. B. and Mark Pickup (2017). Stochastic actor oriented models for network dynamics. ISBN: 9780190228217.

Snijders, Tom A. B., Ruth Ripley, Christian Steglich, Johan Koskinen, Nynke Niezink, Viviana Amati, Christoph Stadtfeld, James Hollway (IHEID), Per Block, Robert Krause, Charlotte Greenan, Josh Lospinoso, Michael Schweinberger, Mark Huisman, Krists Boitmanis, Felix Schoenenberger, Mark Ortmann, Marion Hoffman, Robert Hellpap, and Alvaro Uzaheta (2021). RSiena: Siena - Simulation Investigation for Empirical Network Analysis.

Snijders, Tom A. B., Gerhard G. van de Bunt, and Christian E. G. Steglich (2010). Introduction to stochastic actor-based models for network dynamics. Social Networks 32(1), 44–60.

Sobel, Michael (2006). What do randomized studies of housing mobility demonstrate?: causal inference in the face of interference. Journal of the American Statistical Association 101(476), 1398–1407.

Splawa-Neyman, Jerzy, D. M. Dabrowska, and T. P. Speed (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science 5(4), 465–472. Publisher: Institute of Mathematical Statistics.

Strauss, David and Michael Ikeda (1990). Pseudolikelihood estimation for social networks. Journal of the American Statistical Association 85(409), 204–212.

Swanson, Sonja A., Miguel A. Hernán, Matthew Miller, James M. Robins, and Thomas S. Richardson (2018). Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. Journal of the American Statistical Association 113(522), 933–947.

Tchetgen Tchetgen, Eric J. and Tyler J. VanderWeele (2012). On causal inference in the presence of interference. Statistical Methods in Medical Research 21(1), 55–75.

Thiemichen, Stephanie, Nial Friel, Alberto Caimo, and Gšran Kauermann (2015). Bayesian exponential random graph models with nodal random effects. arXiv:1407.6895 [stat]. arXiv: 1407.6895.

Thiemichen, S., N. Friel, A. Caimo, and G. Kauermann (2016). Bayesian exponential random graph models with nodal random effects. Social Networks 46, 11–28.

Toulis, Panos and Edward Kao (2013). Estimation of causal peer influence effects. Journal of Machine Learning Research 28, 9.

van der Laan, Mark J. (2014). Causal inference for a population of causally connected units. Journal of Causal Inference 2(1), 13–74.

Victor, Jennifer Nicoll, Alexander H. Montgomery, and Mark Lubell (2017). The Oxford Handbook of Political Networks. Oxford: Oxford University Press.

Wasserman, Larry (2004). The bootstrap. In Larry Wasserman (Ed.), All of statistics: a concise course in statistical inference, Springer Texts in Statistics, pp. 107–118. New York, NY: Springer.

Wasserman, Larry (2010). All of Statistics: A Concise Course in Statistical Inference. New York Berlin Heidelberg: Springer.

Winkler, Anderson M., Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols (2014). Permutation inference for the general linear model. NeuroImage 92, 381–397.

Xu, Ran (2020). Statistical methods for the estimation of contagion effects in human disease and health networks. Computational and Structural Biotechnology Journal 18, 1754–1760.

Yan, Ting, Binyan Jiang, Stephen E. Fienberg, and Chenlei Leng (2019). Statistical inference in a directed network model with covariates. Journal of the American Statistical Association 114(526), 857–868.

Young, G. A. and R. L. Smith (2010). Essentials of statistical inference (1 edition ed.). Cambridge, UK: Cambridge University Press.

Appendix A

Appendix for Causal estimation of spillover effects in a social network setting: increasing confidence in positive sexual health attitudes

A.1 Summary statistics for simulated networks

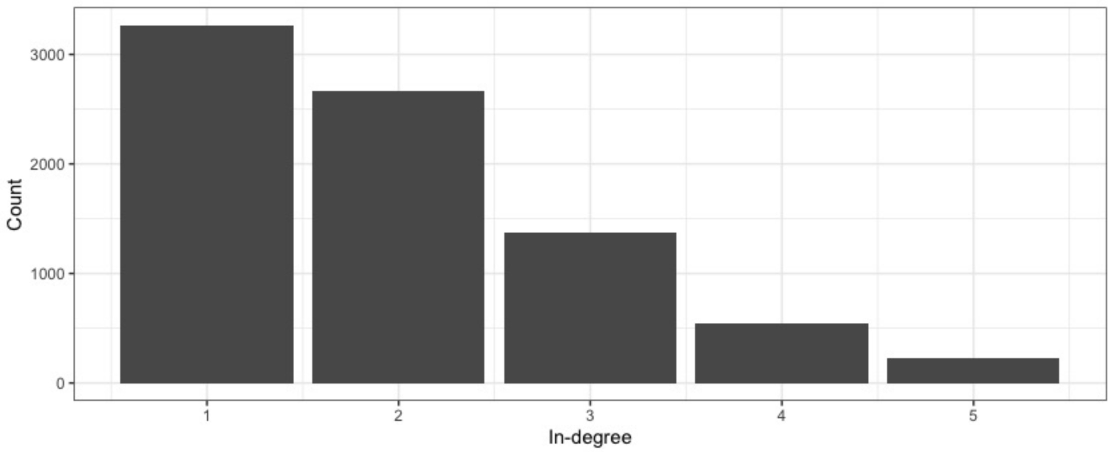


Figure A-1: In-degree distribution for simulation results.

Table A.1: Summary statistics for 100 simulated schools

Variable: $x^{ind}_{1\,i}$	Average value across all schools	Median value across all schools
0	2	2
1	2	2
Variable: $x^{ind}_{2\,i}$	Average value across all schools	Median value across all schools
0	2.02	2
1	2.00	2
2	1.97	2
3	2.02	2
4	2.03	2
5	1.89	2
6	2.06	2
7	1.69	2
8	1.89	1
Variable: In-degree	Average share of treated peers	Average number of treated peers
1	0.15	0.152
2	0.14	0.339
3	0.18	0.496
4	0.13	0.628
5	0.17	0.851

A.2 No-interference Scenario

Table A.2: Bias and standard error (SE) in estimation of direct effect comparing 100 vs 5 schools across 100 simulation replicates. True value of direct effect: 100 schools = 10.49; 5 schools = 10.49

Estimation method	Bias		Standard Deviation			
			Bootstrap SE		Empirical SD	
	Simulation setting		Simulation setting		Simulation setting	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	-4.97	-5.21	0.24	1.03	0.26	1.01
Adjusted	-4.57	-4.88	0.18	0.80	0.23	0.94
Splines (correct)	-0.54	-1.05	0.42	1.86	0.40	1.85
Splines (incorrect PS)	1.28	0.95	0.40	1.72	0.34	1.86
FAM (correct)	-0.50	-0.51	0.20	0.78	0.09	0.34
FAM (incorrect outcome)	0.75	0.65	0.30	1.34	0.16	0.71
FAM (incorrect PS)	-2.20	-2.25	0.15	0.83	0.20	0.89

Table A.3: Bias and standard error (SE) in estimation of spillover effect when ($Z = 0$) comparing 100 vs 5 schools across 100 simulation replicates. True value of direct effect: 100 schools = 1.70; 5 schools = 1.71

Estimation method	Bias		SE			
			Average Bootstrap SE		Monte Carlo SE	
	Simulation setting		Simulation setting		Simulation setting	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	1.93	1.94	0.07	0.33	0.12	0.44
Adjusted	-0.99	-1.01	0.03	0.15	0.03	0.14
Splines (correct)	0.15	0.22	0.10	0.40	0.13	0.48
Splines (incorrect PS)	0.15	0.21	0.10	0.40	0.13	0.48
FAM (correct)	-0.06	-0.04	0.04	0.30	0.03	0.12
FAM (incorrect outcome)	3.18	3.30	0.11	0.59	0.15	0.69
FAM (incorrect PS)	-0.11	0.01	0.06	0.28	0.03	0.33

Table A.4: Bias and standard error (SE) in estimation of spillover effect when ($Z = 1$) comparing 100 vs 5 schools across 100 simulation replicates. True value of direct effect: 100 schools = 1.70; 5 schools = 1.70

Estimation method	Bias		SE			
			Average Bootstrap SE		Monte Carlo SE	
	Simulation setting		Simulation setting		Simulation setting	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	1.45	1.60	0.11	0.48	0.10	0.57
Adjusted	-1.53	-1.53	0.06	0.30	0.06	0.30
Splines (correct)	0.52	0.59	0.13	0.60	0.14	0.66
Splines (incorrect PS)	0.01	0.10	0.12	0.55	0.13	0.58
FAM (correct)	-0.06	-0.08	0.09	0.39	0.06	0.28
FAM (incorrect outcome)	3.35	3.31	0.21	0.95	0.22	0.96
FAM (incorrect PS)	-0.50	-0.52	0.07	0.37	0.09	0.40

Table A.5: Nominal 95% coverage by Wald-type confidence intervals with standard errors based on 500 bootstrap resamples

Estimation method	Main effect		Spillover effect ($Z = 0$)		Spillover effect ($Z = 1$)	
	Simulation setting		Simulation setting		Simulation setting	
	100 schools	5 schools	100 schools	5 schools	100 schools	5 schools
Unadjusted	0.00	0.00	0.00	0.00	0.00	0.10
Adjusted	0.00	0.00	0.00	0.00	0.00	0.00
Splines (correct)	0.76	0.92	0.64	0.90	0.02	0.88
Splines (incorrect PS)	0.08	0.86	0.68	0.92	0.96	0.94
FAM (correct)	0.22	1.00	0.80	1.00	1.00	1.00
FAM (incorrect outcome)	0.18	1.00	0.00	0.00	0.00	0.06
FAM (incorrect PS)	0.00	0.18	0.48	0.98	0.00	0.76

A.3 Additional STASH information

Figure A-2 shows the distribution of this outcome measure for all students in the 6 schools. Figure A-3 shows this distribution for the two relevant statistics that we are considering: in-degree and out-degree.

Figure A-2: Confidence in Sexual Health (CSH) Index: Sample of 605 students - STASH - 2018

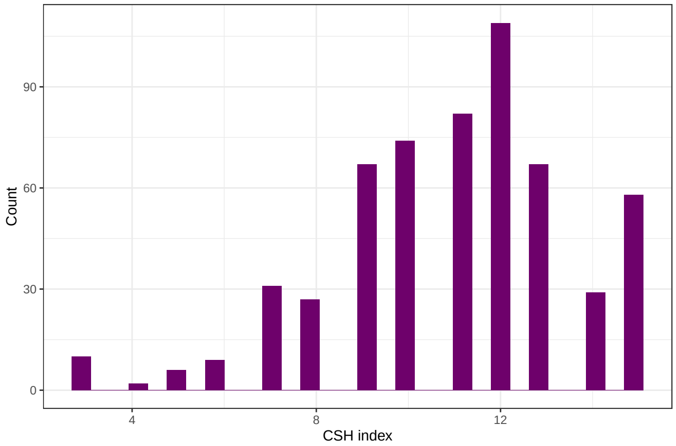


Figure A-3: Degree distribution: Sample of 605 students - STASH - 2018

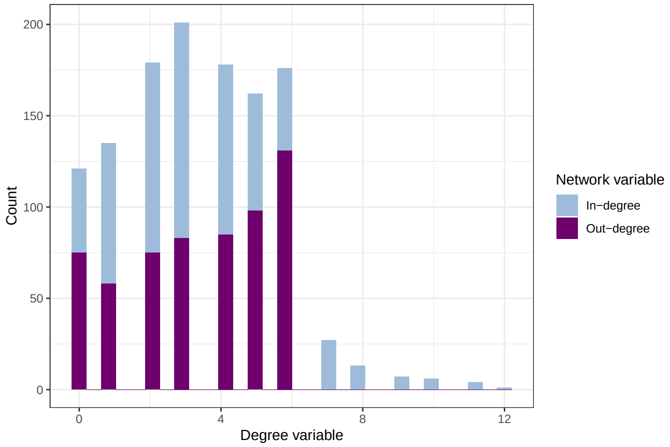


Table A.6: Summary statistics for STASH feasibility study - 605 students - 2018

	Control	Exposed to PS	PS	PS and exposed to PS	Overall
	(N=360)	(N=163)	(N=21)	(N=61)	(N=605)
Gender (assigned at birth)					
Female	208 (57.8%)	93 (57.1%)	10 (47.6%)	36 (59.0%)	347 (57.4%)
Male	151 (41.9%)	70 (42.9%)	11 (52.4%)	23 (37.7%)	255 (42.1%)
Missing	1 (0.3%)	0 (0%)	0 (0%)	2 (3.3%)	3 (0.5%)
How many peers from your school do you think have had sexual intercourse?					
A third	107 (29.7%)	64 (39.3%)	8 (38.1%)	30 (49.2%)	209 (34.5%)
Few	67 (18.6%)	16 (9.8%)	2 (9.5%)	2 (3.3%)	87 (14.4%)
Half	85 (23.6%)	43 (26.4%)	6 (28.6%)	14 (23.0%)	148 (24.5%)
Most	42 (11.7%)	19 (11.7%)	2 (9.5%)	8 (13.1%)	71 (11.7%)
Two thirds	48 (13.3%)	18 (11.0%)	3 (14.3%)	6 (9.8%)	75 (12.4%)
Missing	11 (3.1%)	3 (1.8%)	0 (0%)	1 (1.6%)	15 (2.5%)
Do you talk to your friends about your body?					
No	162 (45.0%)	70 (42.9%)	7 (33.3%)	17 (27.9%)	256 (42.3%)
Yes	183 (50.8%)	91 (55.8%)	14 (66.7%)	42 (68.9%)	330 (54.5%)
Missing	15 (4.2%)	2 (1.2%)	0 (0%)	2 (3.3%)	19 (3.1%)
Do you talk to your friends about STIs?					
No	259 (71.9%)	132 (81.0%)	15 (71.4%)	28 (45.9%)	434 (71.7%)
Yes	80 (22.2%)	29 (17.8%)	6 (28.6%)	31 (50.8%)	146 (24.1%)
Missing	21 (5.8%)	2 (1.2%)	0 (0%)	2 (3.3%)	25 (4.1%)
Have you had sexual intercourse?					
No	256 (71.1%)	127 (77.9%)	20 (95.2%)	42 (68.9%)	445 (73.6%)
Yes	70 (19.4%)	33 (20.2%)	1 (4.8%)	16 (26.2%)	120 (19.8%)
Missing	34 (9.4%)	3 (1.8%)	0 (0%)	3 (4.9%)	40 (6.6%)
Did you talk to someone about STASH?					
No	232 (64.4%)	96 (58.9%)	0 (0%)	0 (0%)	328 (54.2%)
Yes	52 (14.4%)	48 (29.4%)	0 (0%)	0 (0%)	100 (16.5%)
Missing	76 (21.1%)	19 (11.7%)	21 (100%)	61 (100%)	177 (29.3%)
Did you ask a peer supporter a question about sex?					
No	256 (71.1%)	123 (75.5%)	0 (0%)	0 (0%)	379 (62.6%)
Yes	21 (5.8%)	18 (11.0%)	0 (0%)	0 (0%)	39 (6.4%)
Missing	83 (23.1%)	22 (13.5%)	21 (100%)	61 (100%)	187 (30.9%)
Level of self-esteem [a]					
Mean (SD)	2.99 (1.19)	2.86 (1.28)	3.05 (1.32)	3.25 (1.17)	2.98 (1.22)
Median [Min, Max]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]
Missing	17 (4.7%)	3 (1.8%)	0 (0%)	2 (3.3%)	22 (3.6%)
Confidence in Sexual Health index					
Mean (SD)	10.8 (2.56)	10.8 (2.68)	11.3 (2.97)	12.3 (1.96)	11.0 (2.59)
Median [Min, Max]	11.0 [3.00, 15.0]	11.0 [3.00, 15.0]	12.0 [3.00, 15.0]	12.0 [7.00, 15.0]	11.0 [3.00, 15.0]
Missing	27 (7.5%)	3 (1.8%)	0 (0%)	4 (6.6%)	34 (5.6%)

[a] (from 1 to 5), where one is the lowest