Causal Mechanism with Interference

Siwon Ryu*

Abstract

This study examines identification and estimation in the presence of social interactions. Potential outcome depends on both an individual's own treatment and the exposures generated by others' treatment statuses and the underlying network structure. When the network itself is impacted by treatment, the individual's treatment can also alter these exposures. In such cases, the treatment effect can be decomposed into direct and indirect components. The direct effect represents the causal impact on the outcome when exposures are fixed, while the indirect effect captures the influence of changes in the exposure distribution due to the individual's own treatment. As a result, observed exposures can be viewed as mediators of the individual's treatment effect. I employ a causal mediation framework to identify and decompose these effects and propose an estimation method. The performance of the estimator is then evaluated through Monte Carlo simulations, followed by an empirical application examining the impact of coeducational high schools on academic performance.

Keywords: Causal inference; Network change; Mediation Effects

*Department of Economics, Rice University. Email: Siwon.Ryu@rice.edu

1 Introduction

The identification and estimation of causal effects of a program or policy are of significant interest in economic analysis. Rubin's causal model is a widely used approach for addressing causal effects (e.g., Rubin (1974), Imbens and Rubin (2010)). A key assumption in this framework is the Stable Unit Treatment Value Assumption (SUTVA), which assumes that each individual's potential outcomes are solely determined by their own treatment status. However, as highlighted by Kline and Tamer (2020), social interactions offer an additional mechanism through which a program may influence outcomes. In such cases, potential outcomes can depend on both an individual's own treatment status and the treatment status of others within their network.

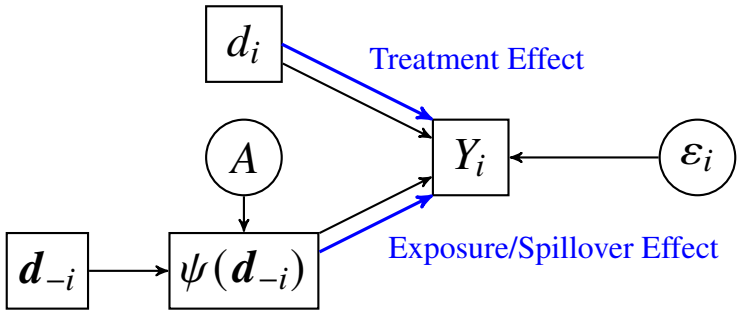
Consider a society composed of N individuals where each individual interacts with others within their neighborhood. The potential outcome for an individual i can be expressed as $Y_i(d_i, \boldsymbol{d}_{-i}) = m(d_i, \boldsymbol{d}_{-i}, \varepsilon_i)$, with some response function $m(\cdot)$, where the first argument $d \in \{0,1\}$ represents the individual's own treatment assignment, the second argument $\boldsymbol{d}_{-i} \in \{0,1\}^{N-1}$ is a vector representing the treatment assignments of the all individuals in the society other than i, and the last argument ε_i is an individual error term. This model violates SUTVA if $Y_i(d_i, \boldsymbol{d}_{-i}) \neq Y_i(d_i, \boldsymbol{d}'_{-i})$ for some $\boldsymbol{d}_{-i} \neq \boldsymbol{d}'_{-i}$. Additionally, note that each individual has 2^N potential outcomes based on the possible number of treatment statuses. Therefore, the number of potential outcomes expands exponentially in the number of individuals. This introduces challenges in defining and identifying the treatment effects.

However, the number of *effective* treatment is likely to be much smaller 2^N. For example, the potential outcome may depend on the individual's own treatment status and the number of treated friends. If each individual has $M < N$ friends, there are $2M$ possible treatment situations, which is considerably fewer than 2^N. Here, the number of treated friends is often called an exposure. In general, if there exists a function $\psi : \{0,1\}^{N-1} \to \Psi \subset \mathbb{R}^K$ for some K, such that $\psi(\boldsymbol{d}_{-i}) = \psi(\boldsymbol{d}'_{-i})$ implies $Y_i(d_i, \boldsymbol{d}_{-i}) = Y_i(d_i, \boldsymbol{d}'_{-i})$ with probability 1, then ψ is called an exposure map, or treatment rule in the literature.

In this setting, causal effects can be defined as follows. First, the treatment effect is the impact of changes in an individual's own treatment status on the outcome when exposures are fixed. Second, spillover or exposure effects refer to the impact of changes in exposures on the outcome when the individual's own treatment status remains fixed. This situation is

described in Figure 1.

Figure 1: Direct and indirect effects



Recent studies emphasize the importance of spillover or exposure effects in program evaluations, as economic agents typically interact with others. As shown in Figure 1, the exposure $\psi_i(\boldsymbol{d}_{-i})$ generally depends on not only the others' treatment $\boldsymbol{d}_{-i}$, but the underlying network structure, represented by the $N \times N$ adjacency matrix $\boldsymbol{A}$. To account for the network, the exposure map can be redefined as $\psi_i : \{0,1\}^{N-1} \times \mathcal{A} \to \Psi$, where $\mathcal{A}$ represents the space of all possible networks with N nodes. For example, the exposure consists of the number of treated friends. The spillover effect then refers to the causal effect of changes in others' treatments via changes in exposure, while holding the network fixed or independent of the treatments.

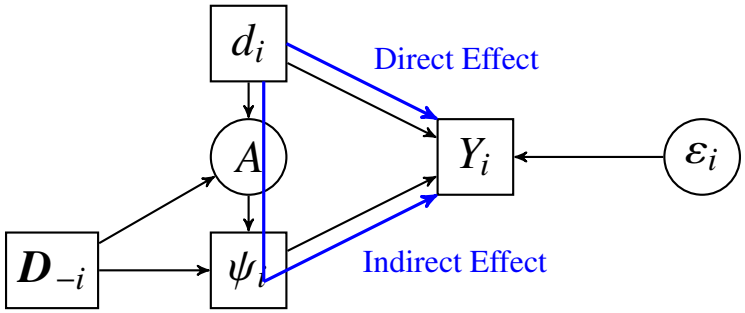
However, empirical evidence suggests that the underlying network may also be affected by the treatment program (e.g., Dupas, Keats, and Robinson (2019), Comola and Prina (2021), Banerjee et al. (2024)). When the network structure is influenced by the treatment, the exposure depends indirectly on each individual's own treatment as well, because the potential exposure is written as $\psi(\boldsymbol{d}_{-i}, \boldsymbol{A}(d_i, \boldsymbol{d}_{-i}))$. Let D_i and $\boldsymbol{D}_{-i}$ represent the treatment indicator for individual i and the treatment vector for others, respectively. With a slight abuse of notation, we can define the counterfactual exposure for individual i as $\psi_i(d_i) = \psi(\boldsymbol{D}_{-i}, \boldsymbol{A}(d_i, \boldsymbol{D}_{-i}))$. The causal effect on the outcome can then be decomposed as:

$$Y_i(1,\psi_i(1)) - Y_i(0,\psi_i(0)) = Y_i(1,\psi_i(0)) - Y_i(0,\psi_i(0)) + Y_i(1,\psi_i(1)) - Y_i(1,\psi_i(0)).$$

The first term on the right-hand side represents the direct effect, which captures the impact of treatment when exposure remains unchanged. The second term represents the indirect effect, reflecting the effect of changes in exposure resulting from the individual's own treatment status. For example, if the exposure is defined as the number of treated friends, the direct

effect measures the casual effect of the treatment when the number of treated friends are fixed. In contrast, the indirect effect captures the difference between potential outcomes when only the number of treated friends changes due to the own treatment. This situation is described in Figure 2.

Figure 2: Direct and indirect effects



The primary contribution of this paper is the separate identification and estimation of direct and indirect treatment effects using a causal mediation model. By decomposing treatment effects, we can better understand the mechanisms through which a policy/program influence outcomes. If the dataset contains complete information on the network adjacency matrix, detailed exposure maps can be constructed. However, this framework is also applicable in situations where full network information is unavailable, as the model only requires data on observed outcomes, treatment status, and some exposures, making it highly advantageous in many of scenarios.

To identify the decomposition, I use a mediation model where exposure serves as a mediator for the treatment effects on potential outcomes. I propose corresponding frequency estimators and derive their asymptotic properties. Since exposures are influenced by others' treatments and the underlying network structure, they are not independent across individuals. Consequently, the asymptotic theory is based on a boundedness assumption in the data dependency graph, which is compatible with either a sparse single large network or a network consisting of multiple independent groups.

The outline of this paper is as follows. Section 2, the settings and model used in this paper are introduced. Section 3 discusses identification and estimation. Section 4 shows results from Monte Carlo studies to verify the performance of estimators. Section 5 is an empirical application to illustrate the proposed method. Section 6 concludes.

Related Literature

The Rubin causal model (e.g., Rubin (1974), Imbens and Rubin (2010)), based on the potential outcome framework, is widely used in economic analysis for identifying and estimating treatment effects. This model assumes SUTVA, which excludes interactions between individuals affecting their potential outcomes. In this context, Manski (2013) refers to SUTVA as the individualistic treatment response (ITR). However, in the presence of social interactions, an individual's treatment response may depend on the entire treatment vector within their society. Manski (2013) identifies the distribution of potential outcomes under the constant treatment response (CTR) assumption. Let $Y_i(\boldsymbol{d})$ represent the potential outcome for individual i when they face the treatment vector $\boldsymbol{d}$. The CTR assumption states that there exists a function c_i such that $c_i(\boldsymbol{d}) = c_i(\boldsymbol{d}')$ implies $Y_i(\boldsymbol{d}) = Y_i(\boldsymbol{d}')$. Manski refers to the image of this function as the set of effective treatments. The function c_i is called the exposure map or treatment rule in the literature. The CTR assumption generalizes the ITR or SUTVA, as these are special cases where $c_i(\boldsymbol{d}) = d_i$, and c_i is the identity function for all i.

Because analyzing the general unrestricted model, where $c_i(\boldsymbol{d}) = \boldsymbol{d}$ is challenging due to the dimensionality of the potential outcome space, the concept of effective treatment is commonly used in the literature. Studies often assume the effective treatment is finite-dimensional, reducing the complexity of both the treatment and the number of potential outcomes. Forastiere, Airoldi, and Mealli (2021) propose the Stable Unit Treatment Value on Neighborhood Assumption (SUTNVA), where potential outcomes depend only on the treatment vector of the individual's neighborhood. This is a natural extension of SUTVA to settings involving neighborhood interactions. Aronow and Samii (2017), Vazquez-Bare (2023) assume the neighborhood size is fixed, meaning each individual has a finite number of potential outcomes.

As such studies, once we fix the dimension of effective treatment, or the domain of potential outcomes, then identifying treatment effect parameters are similar to those of multiple treatment model. Corresponding independence assumptions or ignorabilities imply the identification. As related works, Leung (2020) focus on the number of treated neighbors and the number of neighbors (i.e., degree) as the exposures. The model is therefore a linear-in-means model without endogenous peer effect. The author identifies the average and quantile treatment, exposure effects, and derive their asymptotic normality by using a new version of asymptotic theories. Vazquez-Bare (2023) derive general identification argument of this type

of model and propose some estimators. The author assume there are independent groups in data, and the interaction arises within group. By controling the size of group and the number of group, he show asymptotic normality of the frequency estimator.

One important problem is the possibility of misspecification of the exposure map. If the exposure map is incorrect, then the identification and corresponding estimators would be misleading. In this respect, Vazquez-Bare (2023) also shows that if the true exposure map is coarser then the exposure map used in the estimator, then the average potential outcomes are identified as the usualy frequcney estimands, and some weighted average of those frequency estimands otherwise. Intuitively, the less coarsity of true exposure map means the information contained in the exposure map is more rich than that from the true exposure map. Leung (2022) assumes *approximately neighborhood interaction (ANI)* assumption to identify the treatment parameteres when the exposure map is possibly misspecified. This assumption is that the potential outcome distributions are primarily determined by the neighbors within some close distance. However, I assume we have correctly specified exposure map in this study, to focus on the dependency between exposure and the treatment assignments.

Most studies in this literature assume the underlying social network is fixed or independent of treatment assignment. These assumptions exclude the possibility that a policy can change the underlying random graphs of social networks. Some studies consider this possibility. Comola and Prina (2021) suggests a two-period model in that treatments are assigned in the first period, and the network can change in the second period. The outcome in each period follows the linear in means model as in Bramoullé, Djebbari, and Fortin (2009), so they use the similar identifying assumption that there is no intransitive triad in networks of both periods. To apply this model, researchers need to have full information about network adjacency matrices in both periods. However, It would sometimes be hard to get such descent information.

If potential outcomes are functions of the own treatment status and the exposures, the exposures are determined by the underlying social networks. In this study, I assume that the distributions of exposures are influenced by their own treatment status. That is, the own treatment status could change the functional form of the exposure map, or change the random graph of social networks. In both cases, exposure can be thought of as a mediator of own treatment on potential outcomes. In the literature on mediation models, the main purpose is to figure out the mechanism of how treatment influences potential outcomes by decompos-

ing the total treatment effect into direct and indirect effects. Suppose $M(d)$ is the potential mediator when own treatment status is given by $d \in \{0,1\}$. Then, the observed mediator is $M = dM(1) + (1-d)M(0)$. The direct treatment effect is the effect of the own treatment status when the mediator is fixed, and the indirect treatment effect is the effect of the own treatment only through changes in the mediator. If both treatment assignment and the distributions of mediators are independent of potential outcomes, then it is straightforward to identify distributions of direct and indirect treatment effects separately. Huber (2014, 2019) suggests sequential ignorability assumptions, which are weaker conditions than the full independence to identify the direct and indirect effects, and I follow these assumptions and identification arguments in this study.

Even if the treatments are randomly assigned, the data of outcomes and exposures could be dependent due to social interactions. This dependency makes it tricky to derive asymptotic properties. In the case of the sum of independent random variables, Esseen's method (Esseen (1945)) is convenient to approximate normal distribution. Vazquez-Bare (2023) uses the Berry-Esseen bound to derive the asymptotic normality by assuming exposures are independent across groups. However, it is difficult to apply Esseens' method to dependent data. Instead, Stein's method (Stein (1972)) is widely used to deal with dependent data. As an example, Chen and Shao (2004) provides a version of the central limit theorem with a bounded maximum degree of the dependency graph. Leung (2020) derived conditions on the moment of dependency graph instead of directly applying Stein's method, but in this study, I assume boundedness of the maximum degree of dependency graph and use Stein's method.

2 Model

2.1 Setting and the exposure map

Consider a society consisting of N individuals, and let $D_i \in \{0,1\}$ be the binary treatment indicator for individual i. For any N-vector $\boldsymbol{V} = (V_1, ..., V_N)$, and for each i, let $\boldsymbol{V}_{-i}$ be the $(N-1)$-vector that excludes V_i from $\boldsymbol{V}$. Thus, the treatment statuses of all individuals in the society can be represented by the N-vector $\boldsymbol{D} = (D_1, ..., D_N)$, which is divided by $(D_i, \boldsymbol{D}_{-i})$, where D_i is individual i's treatment status, and $\boldsymbol{D}_{-i}$ is the treatment status of all other individuals.

Each individual interacts with others through an underlying network structure. Let A_{ij} be

an indicator of whether individuals i and j are friends, where $A_{ii} = 0$ for all i. The network can be represented by an $N \times N$ adjacency matrix $\boldsymbol{A}$ with $[\boldsymbol{A}]_{ij} = A_{ij}$.

Now, consider a potential treatment assignment $\boldsymbol{d} = (d_1, ..., d_N) \in \{0,1\}^N$ for all individuals, and again, denote $(d_i, \boldsymbol{d}_{-i})$ as individual i's treatment, and all other individual's treatment vector, respectively. The potential outcome for each individual needs to be defined by a function of the entire treatment vector $\boldsymbol{d}$ in general. However, I assume the other's treatment vector $\boldsymbol{d}_{-i}$ affect the potential outcome via an exposure map. Specifically, suppose there exists a known function $\psi : \{0,1\}^{N-1} \times \mathcal{A} \to \Psi$, where $\mathcal{A}$ is the space of networks of N nodes, and $\Psi \subset \mathbb{R}^K$ with $K < N$. This function satisfies the condition that the potential outcome for individual i is the same for any two treatment vectors $\boldsymbol{d}$ and $\boldsymbol{d}'$ as long as $d_i = d'_i$, and $\psi(\boldsymbol{d}_{-i}, \boldsymbol{A}) = \psi(\boldsymbol{d}'_{-i}, \boldsymbol{A})$.

This function, ψ, is an exposure map that provides a rule that links the other's treatment to an individual's potential outcome. In this study, I assume the exposure map is correctly specified.[1,2] Thus, the potential outcome for individual i is well defined by their own treatment status d_i and their exposure $s_i = \psi(\boldsymbol{d}_{-i}, \boldsymbol{A})$, i.e., we denote the potential outcome for individual i as $Y_i(d_i, s_i) = Y_i(d_i, \psi(\boldsymbol{d}_{-i}, \boldsymbol{A}))$. For example, Leung (2020) shows that if the network is anonymous and individuals interact with their neighbors within 1 network distance, then the potential outcome is determined by an individual's own treatment status d_i, the number of neighbors $\sum_j A_{ij}$, and the number of treated neighbors $\sum_j A_{ij} D_j$. In this case, the exposure map is given by $\psi(\boldsymbol{d}_{-i}, \boldsymbol{A}) = (\sum_j A_{ij}, \sum_j A_{ij} d_j)$.

The treatment effect is defined as the effect of changing an individual's own treatment status on the outcome, i.e., $Y_i(1, s) - Y_i(0, s)$, for a given exposure level s. The spillover or exposure effect is defined as the effect of a change in the exposure on the outcome, i.e., $Y_i(d, s') - Y_i(d, s)$, for some d, s', s. For example, Leung (2022) uses the exposure map $\psi(\boldsymbol{d}_{-i}, \boldsymbol{A}) = \mathbb{1}\left\{\sum_j A_{ij} d_j > 0\right\}$. In this case, the exposure effect represents the difference in potential outcomes between an individual with at least one treated friend and one with no treated friends. The treatment effect is interpreted as the usual causal effect of an individual's own treatment, while the exposure effect reflects the causal effect of changes in the treatment status of others, provided that the underlying network remains fixed.

[1] Aronow and Samii (2017) discuss about when the exposure map is misspecified. Leung (2022) proposed a solution when the exposure map is misspecified by using the concept of *approximated neighborhood inference*.

[2] Constructing a model using an exposure map is convenient when only limited information of the network structure is available in data instead of the full information of the adjacency matrix.

Recent empirical studies suggest that the network can also be influenced by the program. For instance, Comola and Prina (2021) use experimental data from Nepal and find that providing savings accounts to households leads to changes in their network connections. Similarly, Dupas, Keats, and Robinson (2019) use experimental data from Kenya, where households were given free savings accounts. They observe that these households became less dependent on distant family members and more supportive of neighbors and friends within their village. Given this evidence, I assume that the network can also be altered by the treatment, in contrast to previous studies in the literature that assume a fixed or exogenous network.

To account for changes in the network due to the treatment, let $\boldsymbol{A}(\boldsymbol{d}) = \boldsymbol{A}(d_i, \boldsymbol{d}_{-i})$ represent the potential network based on the treatment assignment $\boldsymbol{d} \in \{0,1\}^N$ of all individuals. The exposure for individual i is then given by $\psi(\boldsymbol{d}_{-i}, \boldsymbol{A}(d_i, \boldsymbol{d}_{-i}))$. As a result, changes in exposure now reflect both the treatment status of others and changes in individual i's own treatment status. This suggests that the previously defined exposure effect not only captures the impact of others' treatment status but also includes changes in the network due to individual i's treatment. Consequently, the latter part of the exposure effect needs to be interpreted as an *indirect* treatment effect via changing the exposure level. This situation is described in Figure 2.

To focus on this additional treatment effect, by an abuse of notation, define $\psi_i(d) := \psi(\boldsymbol{D}_{-i}, A(d, \boldsymbol{D}_{-i}))$ for $d \in \{0,1\}$ as the potential exposure for individual i. The observed exposure for individual i is given by $\psi_i = \psi_i(D_i)$. Similarly, abusing notation again, we can define the potential outcome as $Y_i(d, d') := Y_i(d, \psi_i(d'))$, for $d, d' \in \{0,1\}$. The observed outcome is then $Y_i = Y_i(D_i, D_i)$, and therefore $Y_i(1,1)$ or $Y_i(0,0)$ is observed in the sample, while $Y_i(d, d')$ for $d \neq d'$ is never observed. can be viewed as a mediation model, as studied in Huber (2014, 2019), where the exposure $\psi_i(d)$ plays the role of a mediator.

The different distributions of $\psi_i(1)$ and $\psi_i(0)$ are important for the identification. However, if the distributions of $\psi_i(0)$ and $\psi_i(1)$ are identical,[3], the model simplifies to a potential outcomes framework with multiple treatments. In such a case, as discussed in Vazquez-Bare (2023) and Leung (2020), identification of treatment and exposure effects follows from standard independence assumptions. Furthermore, if the exposure map is a constant function, the model reduces to the classical causal model with SUTVA. Therefore, the framework is a generalization of existing methods.

[3] Or, if network is unaffected by the treatment.

2.2 Parameters of interest

This section defines the key parameters of interest in this paper. The potential outcome for an individual i is denoted as $Y_i(d, \psi_i(d))$. The *average overall treatment effect (ATE)* is defined as the mean difference between potential outcomes when the individual's own treatment is exogenously changed:

$$\Delta \equiv E[Y_i(1, \psi_i(1)) - Y_i(0, \psi_i(0))].$$

The ATE captures the average causal effect of the individual's own treatment, including both the direct effect of the treatment and the indirect effect through exposure. As discussed earlier, exposure can be thought of as a mediator of the treatment. Thus, drawing from the literature on causal mediation effects (e.g., Huber (2014, 2019)), the total treatment effect can be decomposed into direct and indirect effects as follows.

The *average direct treatment effect (ADTE)* is defined as the average difference between potential outcomes when the individual's own treatment is exogenously changed, while the mediator (network structure) is held fixed at its potential distribution for a given $d \in \{0, 1\}$:

$$\theta(d) \equiv E[Y_i(1, \psi_i(d)) - Y_i(0, \psi_i(d))].$$

Similarly, *the average indirect treatment effect (AITE)* is defined as the average difference between potential outcomes when the mediator's distribution is exogenously changed, while the individual's own treatment is fixed at $d \in \{0, 1\}$:

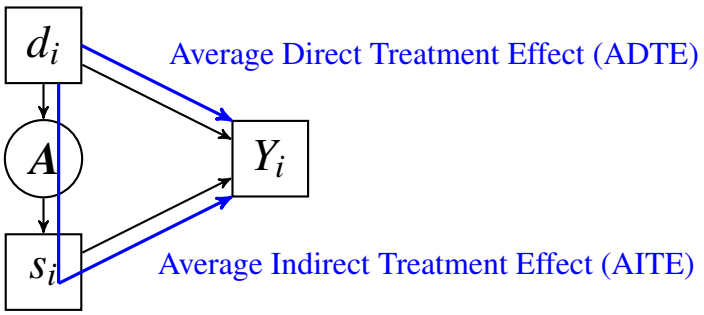
$$\delta(d) \equiv E[Y_i(d, \psi_i(1)) - Y_i(d, \psi_i(0))].$$

Therefore, by construction, the ATE can be decomposed into the sum of the ADTE and AITE: $\Delta = \theta(0) + \delta(1) = \theta(1) + \delta(0)$.

For a more detailed interpretation, consider the potential outcome is determined by a response function h, such that $Y_i(d, s) = h(d, s, \varepsilon_i)$, where $\varepsilon_i \sim F_{\varepsilon_i}$ is the individual-specific error that is independent of both the treatment and the underlying network. Moreover, let $F_{\boldsymbol{D}_{-i}}(\boldsymbol{d}_{-i}) := \Pr(\boldsymbol{D}_{-i} = \boldsymbol{d}_{-i})$ represent the distribution of other's treatment statuses, and for $d \in \{0, 1\}$, let $F_{\boldsymbol{A},d}(\boldsymbol{a}|\boldsymbol{d}_{-i}) := \Pr(\boldsymbol{A}(d, \boldsymbol{D}_{-i}) = \boldsymbol{a}|\boldsymbol{D}_{-i} = \boldsymbol{d}_{-i})$ denote the distribution of potential network links.

As described in Figure 2, the potential outcome $Y_i(d, \psi_i(d))$ is determined by the distributions $(F_{\varepsilon_i}(\cdot), F_{\boldsymbol{D}_{-i}}(\cdot), F_{\boldsymbol{A},1}(\cdot|\cdot), F_{\boldsymbol{A},0}(\cdot|\cdot))$, and therefore, ADTE and AITE measure the impact of one's own treatment, after integrating out those distributions. It can be simply described in Figure 3.

Figure 3: Average direct and indirect effects



Specifically, the ADTE can be written by:

$$
\begin{aligned}
&E[Y_i(\mathbf{1}, d) - Y_i(\mathbf{0}, d)] \\
&= \int \left[\int \{h(\mathbf{1}, \psi(\boldsymbol{d}_{-i}, \boldsymbol{a}), e) - h(\mathbf{0}, \psi(\boldsymbol{d}_{-i}, \boldsymbol{a}), e)\} \underline{dF_{\boldsymbol{A},d}(\boldsymbol{a}|\boldsymbol{d}_{-i})} \right] dF_{\boldsymbol{D}_{-i}}(\boldsymbol{d}_{-i}) dF_{\varepsilon_i}(e).
\end{aligned}
$$

The difference between the inner expectations reflects the variation in potential outcomes due to changes in treatment status, while the network distribution remains fixed at $F_{\boldsymbol{A},d}(\boldsymbol{a}|\boldsymbol{d}_{-i})$. Similarly, the AITE is given by:

$$
\begin{aligned}
&E[Y_i(d, \mathbf{1}) - Y_i(d, \mathbf{0})] \\
&= \int \left[\int h(d, \psi(\boldsymbol{d}_{-i}, \boldsymbol{a}), e) d \left(\underline{F_{\boldsymbol{A},\mathbf{1}}(\boldsymbol{a}|\boldsymbol{d}_{-i})} - \underline{F_{\boldsymbol{A},\mathbf{0}}(\boldsymbol{a}|\boldsymbol{d}_{-i})} \right) \right] dF_{\boldsymbol{D}_{-i}}(\boldsymbol{d}_{-i}) dF_{\varepsilon_i}(e).
\end{aligned}
$$

Here, the difference captures the change in the distribution of the network, from $F_{\boldsymbol{A},0}(\cdot|\boldsymbol{d}_{-i})$ to $F_{\boldsymbol{A},1}(\cdot|\boldsymbol{d}_{-i})$. Therefore, the AITE can be interpreted as the average causal effect of changes in the network on the potential outcome.

Spillover effects, or exposure effects, refer to the impact on potential outcomes when the exposure level is exogenously changed as usual definition in the literature. Let s and s' represent two different values of $\psi_i(d)$. The *exposure effect* is then defined as: $\tau(d, s, s') \equiv E[Y_i(d, s) - Y_i(d, s')]$.

3 Identification and Estimation

3.1 Identification

In this section, we discuss the identification of the parameters ATE, AITE, and ADTE as defined in Section 2.2. To begin, assume we observe a random sample of $\{(Y_i, D_i, \psi_i) : 1 \leq i \leq N\}$. The observed outcome can be expressed as:

$$Y_i = Y_i(D_i, \psi_i(D_i)) = \sum_{d \in \{0,1\}} \sum_{s \in \Psi} \mathbb{1}\{D_i = d, \psi_i = s\} Y_i(d, s) \tag{1}$$

As discussed earlier, either $Y_i(1, \psi_i(1))$ or $Y_i(0, \psi_i(0))$ is observed, but both $Y_i(1, \psi_i(0))$ and $Y_i(0, \psi_i(1))$ are never observed. However, if the treatment is exogenous, we can identify the distributions of the potential exposures $\psi_i(1), \psi_i(0)$. This allows us to identify the average counterfactual outcomes by integrating $Y_i(d, \psi_i(d'))$ over the distribution of $\psi_i(d')$. The following are the identifying assumptions based on Huber (2014) for identifying causal mediation effects.

Assumption 1. $\{Y_i(1, s), Y_i(0, s), \psi_i(1), \psi_i(0) : s \in \Psi\}$ *are independent of* D_i.

Assumption 2. $\{Y_i(1, s), Y_i(0, s) : s \in \Psi\}$ *are independent of* ψ_i *conditional on* D_i.

These assumptions are referred to as *sequential independence*. Note that Assumption 1 states that the potential outcome and potential exposure are independent of the treatment. The distribution of the potential outcome, given the individual's own treatment and exposure, is denoted by (d, s) and comes from the distribution of the unobserved individual error term ε_i. The potential exposure $\psi_i(d)$ of individual i's arises from the distribution of others' treatment $\boldsymbol{D}_{-i}$ and the potential network $\boldsymbol{A}(d, \boldsymbol{D}_{-i})$. Therefore, if the treatment is randomly assigned or exogenously given, i.e., $(\boldsymbol{Y}_i(d, s), \boldsymbol{A}(\boldsymbol{d})') \perp \boldsymbol{D}$ for all (d, s) and $\boldsymbol{d}'$, Assumption 1 is satisfied.

Assumption 2 requires independence between the potential outcome and potential exposure. The randomness of $Y_i(d, s)$ is determined by the distribution of individual error ε_i, while $\psi_i(d)$ is determined by others' treatment vectors and the potential network. Therefore, if the treatment is randomly assigned, this assumption implies independence between ε_i and

the potential network $A(d, \boldsymbol{D}_{-i})$.[4]

Assumption 3. *For each* $d \in \{0,1\}$ *and* $s \in \Psi$, $P(d,s) \equiv P(D_i = d, \psi_i = s) \in (0,1)$.

Assumption 3 is the overlap assumption that ensures the existence of appropriate conditional moments. The following Lemma 1 states that the distributions of interest are identified.

Lemma 1 (Identification of distributions)**.** *Under Assumptions 1, and 2,*

$$G^{d,s}(y) := \Pr\left(Y_i(d,s) \le y\right) = \Pr\left(Y_i \le y | D_i = d, \psi_i = s\right),$$

$$F^{d,d'}(y) := \Pr(Y_i(d, \psi_i(d')) \le y) = \sum_{s \in \Psi} \Pr(Y_i \le y | D_i = d, \psi_i = s) \Pr(\psi_i = s | D_i = d'),$$

Note that the distributions of $Y_i(d,s)$ and $Y_i(d, \psi_i(d))$ are identified in the usual way from the independence assumptions. Identification of $Y(d, \psi_i(d'))$ for $d \neq d'$ requires the support of $\psi_i(1)$ and $\psi_i(0)$ are the same. If $\text{Supp}(\psi_i(1)) \subsetneq \text{Supp}(\psi_i(0))$, then only $F^{0,1}(y)$ is identified, not $F^{1,0}(y)$. Using Lemma 1, we can derive the following result.

Proposition 1 (Identification of Average Effects)**.** *Under Assumptions 1, 2,*

$$\theta(0) = \sum_{s \in \Psi} E\left[Y_i | D_i = 1, \psi_i = s\right] \Pr(\psi_i = s | D_i = 0) - E\left[Y_i | D_i = 0\right],$$

$$\delta(1) = E\left[Y_i | D_i = 1\right] - \sum_{s \in \Psi} E\left[Y_i | D_i = 1, \psi_i = s\right] \Pr(\psi_i = s | D_i = 0),$$

$$\Delta = \delta(1) + \theta(0) = E\left[Y_i | D_i = 1\right] - E\left[Y_i | D_i = 0\right].$$

$\theta(1)$ *and* $\delta(0)$ *are identified similarly with* $\Delta = \theta(1) + \delta(0)$. *For each* $d \in \{0,1\}$ *and for* $s, s' \in \Psi$, *the exposure effects are identified as* $\tau(d,s,s') = E\left[Y_i | D_i = d, \psi_i = s'\right] - E\left[Y_i | D_i = d, \psi_i = s\right]$.

[4] Assumption 2 may fail if a common factor determines both potential outcomes and exposures. This occurs, for example, when there is a random variable X_i with a nontrivial distribution such that $Y_i(\boldsymbol{d}, \boldsymbol{d}_{-i}) = m(d_i, \boldsymbol{d}_{-i}, X_i, \varepsilon_i)$ and $\psi_i = \psi(D_i, \boldsymbol{D}_{-i}, X_i)$. If we observe X_i, then Assumptions 1 and 2 can be stated additionally conditional on X_i.

3.2 Estimation and Inference

Based on the identification results in Proposition 1, we can construct estimators for the averages of potential outcomes and treatment parameters. For notational simplicity, let $\mathbb{1}_i(d,s) = \mathbb{1}\{D_i = d, \psi_i = s\}$, and $\mathbb{1}_i(d) = \mathbb{1}\{D_i = d\}$. Define $N(d,s) \equiv \sum_{i=1}^N \mathbb{1}_i(d,s)$, and $N(d) \equiv \sum_{i=1}^N \mathbb{1}_i(d) = \sum_{s\in\Psi} N(d,s)$ be the potential number of individuals with $D_i = d, \psi_i = s$, and $D_i = d$, respectively. Next, define $\nu(d,s) := E[Y(d,s)] = E[Y_i|D_i = d, \psi_i = s]$ for $d \in \{0,1\}, s \in \Psi$, $\mu(d,d') := E[Y(d,\psi_i(d'))]$ for $d,d' \in \{0,1\}$, and $\mu(d) := \mu(d,d)$ as the average potential outcomes. These average outcomes can be estimated by the following frequency estimators:

$$
\begin{aligned}
\hat{\nu}(d,s) &= \frac{1}{N(d,s)} \sum_{i=1}^N \mathbb{1}_i(d,s) Y_i, \\
\hat{\mu}(d) &= \frac{1}{N(d)} \sum_{i=1}^N \mathbb{1}\{D_i = d\} Y_i, \\
\hat{\mu}(d,d') &= \frac{1}{N(d')} \sum_{j=1}^N \hat{\nu}(d,\psi_j) \mathbb{1}\{D_j = d'\} \\
&= \frac{1}{N(d')} \sum_{s\in\Psi} \sum_{j=1}^N \hat{\nu}(d,s) \mathbb{1}_j(d',s) = \sum_{s\in\Psi} \hat{\nu}(d,s) \frac{N(d',s)}{N(d')}.
\end{aligned}
$$

Here, $\nu(d,s)$ is the sample average of the observed outcomes in the subsample where $D_i = d$, and $\psi_i = s$. Note that these estimators are undefined if the corresponding cells are empty. Similarly, $\hat{\mu}(d)$ is the sample average of outcomes on the subsample with $D_i = d$. The estimator $\hat{\mu}(d,d')$ is the weighted average of the average potential outcomes $Y(d,s)$, where the weights are the sample analog of $\Pr(\psi_i = s|D_i = d')$. Using these estimators, the overall, direct, and indirect treatment effects can be estimated as follows: $\hat{\Delta} = \hat{\mu}(1) - \hat{\mu}(0)$, $\hat{\Delta} = \hat{\mu}(1) - \hat{\mu}(0)$, $\hat{\delta}(d) = \hat{\mu}(d,1) - \hat{\mu}(d,0)$, and $\hat{\tau}(d,s,s') = \hat{\nu}(d,s) - \hat{\nu}(d,s')$.

Next, we derive the consistency and asymptotic normality of estimators for average outcomes. As mentioned earlier, Y_i may exhibit dependence across individuals. However, because the treatment is randomly assigned, the potential outcomes can be independently and identically distributed. Therefore, we assume the following:

Assumption 4. *For $d \in \{0,1\}$ and $s \in \Psi$, (i) $\{Y_i(d,s)\}_i$ are i.i.d.; (ii) $E[Y_i(d,s)^6] < \infty$.*

To apply normal approximation in a dependent data, we need to restrict about the dependency. Specifically, for each individual i, the number of dependent individual grows slower than the number of individuals N. The primary condition is stated in Assumption 5:

Assumption 5. *Let $C_i = (D_i, \psi_i)$, and g_{ij} represent an indicator of whether for two different individuals i and j, C_i is independent of C_j, i.e., $g_{ij} = \mathbb{1}\{i = j, or\ C_i \not\perp C_j\}$. Then, $\sum_{j=1}^{N} g_{ij} = O(N^{\delta})$ for some $0 < \delta < 1$.*

Then, the estimators for average potential outcome is asymptotically normal and the plug-in standard error is asymptotically valid. Proposition 2 summarizes the result.

Proposition 2. *Under Assumptions 1-5, for each $d \in \{0,1\}$ and $s \in \Psi$,*

$$\hat{V}(d,s)^{-1/2}\sqrt{N}(\hat{\nu}(d,s) - \nu(d,s)) \xrightarrow{d} N(0,1),$$
$$\hat{V}_{\mu}(d,d')^{-1/2}\sqrt{N}(\hat{\mu}(d,d') - \mu(d,d')) \xrightarrow{d} N(0,1),$$

where $\hat{V}(d,s)$, $\hat{V}_{\mu}(d,d')$ are estimator for

$$V(d,s) = \frac{\mathrm{Var}(Y_i(d,s))}{P(d,s)},$$
$$V_{\mu}(d,d') = \sum_{s \in \Psi} \frac{P(d',s)^2}{P(d')^2} V(d,s) + \sum_{s \neq s' \in \Psi} \frac{P(d',s)}{P(d')} \frac{P(d',s')}{P(d')} E[Y(d,s)]E[Y(d,s')],$$

respectively, by replacing $\mathrm{Var}(Y(d,s))$ *as* $N(d,s)^{-1} \sum_{i=1}^{N} \mathbb{1}_i(d,s)[Y_i - \hat{\nu}(d,s)]^2$.

4 Simulation

This section evaluate the finite sample performance and the asymptotic result derived in Section 3 using Monte Carlo simulations. Specifically, the mean squared errors and the coverage rates of the estimators provide simulation evidence for asymptotic normality. The data for the simulation consists of N units, where each unit has a binary treatment D_i, drawn from a Bernoulli experiment with probability q. The exposure map for each unit is defined as:

$$\psi_i(d) = (\psi_{i,1}(d), \psi_{i,2}(d)) = (M_i(d), N - M_i(d)),$$

where $M_i(d)$ represents the number of treated neighbors when $D_i = d$. Here, $M_i(d)$ is drawn from a truncated normal distribution on $[0, M]$ with mean $Mp(d)$ and variance σ^2, where $p(d) = p_1^d p_0^{1-d}$. Potential outcomes are generated by $Y_i(d,s) = \theta_1 + \theta_2 d + \theta_3 \psi_{i,1}(d) + \theta_4 \psi_{i,2}(d) + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$. Therefore, this DGP represents a linear-in-sums model without endogenous peer effect. Parameters are set by $\theta = (1,2,3,4)', M = 10, \sigma = 5, p_1 = 0.26298, p_2 = 0.73701, q = 0.5$. The choice of θ, p_1, p_2 makes the true ATE as 6, DTE as 2, and ITE as 4. $\sigma = 5$ is for overlapping assumption. Table 1, and Table 2 show mean squared errors of each estimator, which are defined as $\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_s - \theta^\star)^2$, where $\hat{\theta}_s$ is the estimate in sth replication, $\theta^\star$ is the true parameter value, and the number of replication is $S = 10,000$. This shows that the MSE, hence both bias and variance, of each estimator converges to zero with the theoretical rate of $N^{-1/2}$. Moreover, it presents that the error is small in a relatively small sample size. Next, Table 3, and Table 4 show the coverage rates of each estimator, which are defined as $C(\hat{\theta}, \theta) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\{\theta \in CI(\hat{\theta}_s)\}$, where $CI(\hat{\theta}_s) = [\hat{\theta}_s - 1.96\text{se}(\hat{\theta}), \hat{\theta}_s + 1.96\text{se}(\hat{\theta})]$ is the 95% confidence interval.

Table 1: Mean Squared Errors of Average Potential Outcomes

Design	N	$\mu(0)$	$\mu(0,1)$	$\mu(1,0)$	$\mu(1,1)$
1	500	0.032	0.0325	0.0439	0.0447
	1,000	0.0164	0.0165	0.0176	0.017
	5,000	0.0033	0.0032	0.0033	0.0034
	10,000	0.0016	0.0016	0.0017	0.0017

Table 2: Mean Squared Errors of Treatment Effects

Design	N	Δ	$\theta(1)$	$\theta(0)$	$\delta(1)$	$\delta(0)$
1	500	0.0648	0.0189	0.0201	0.0686	0.0701
	1,000	0.0329	0.0051	0.0047	0.0297	0.0299
	5,000	0.0064	0.0009	0.0009	0.0058	0.0058
	10,000	0.0033	0.0005	0.0005	0.003	0.003

Notes. MSEs are computed by 10,000 simulations. $MSE = \frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_s - \theta)^2$, where θ is the true value of parameters from the design.

Table 3: 95% Coverage Rates of Average Potential Outcomes

Design	N	$\mu(0)$	$\mu(0,1)$	$\mu(1,0)$	$\mu(1,1)$
1	500	0.9692	0.9813	0.9788	0.9721
	1,000	0.9759	0.986	0.9845	0.9744
	5,000	0.9773	0.9843	0.987	0.9765
	10,000	0.9774	0.9841	0.9847	0.9754

Table 4: 95% Coverage Rates of Treatment Effects

Design	N	Δ	$\theta(1)$	$\theta(0)$	$\delta(1)$	$\delta(0)$
1	500	0.9944	0.9959	0.9958	0.9971	0.9976
	1,000	0.9987	0.9999	0.9999	0.9997	0.9998
	5,000	0.9983	1	1	0.9998	0.9996
	10,000	0.9978	1	1	0.9993	0.9994

Notes. Coverage probabilities are computed by 10,000 replications. For the treatment parameters, the confidence intervals are computed by ignoring the asymptotic covariances of average potential outcomes, resulting in a conservative coverage.

5 Empirical Application

This section presents a simple empirical analysis to demonstrate how the decomposition proposed in Section 3 can be applied to real data. To estimate the treatment effects and decompose them, we require data that consists of a random experiment, outcome, and exposure map. In Korea, high school assignments are nearly random when students graduate from middle school. I utilize this random assignment to estimate the impact of attending coeducational high schools on academic performance.

The data used in this application comes from the Korean Education and Employment Panel II (KEEP II), collected by the Korean Research Institute for Vocational Education and Training (KRIVET). The population consists of second-year high school students in 2016. The initial sample includes 10,558 students from 416 schools.

When students graduate from middle school, they select the type of high school they

wish to apply to. After making their choice, high school assignments are nearly random within each type and region, based on the student's residential address. This study leverages the exogenous variation from the random assignment of high schools.

There are five types of high schools in Korea. General high schools are the most common, and most students attending these schools aim to enter a university after graduation. Engineering high schools prepare students for immediate employment upon graduation. Special, Science, and Foreign Language high schools require entrance exams, so students attending these schools are not randomly assigned. In this application, I focus solely on general high schools.

I set two outcomes to assess students' academic performance. The first outcome is the student's relative grade within their school. In each high school, students are ranked across nine grade levels, with grade 1 representing the top 4% and grade 9 representing the bottom 4%.[5] The second outcome is an indicator of whether the student attends a university in Seoul. Since many of the top-ranked universities are located in Seoul, this outcome is intended to capture students' academic achievements.[6]

Table 5 shows the distribution of outcome based on school types and gender. The average relative grade for female students is 3.92 in all-female schools and 3.81 in coeducational schools. For male students, the average relative grade is 4.11 in all-male schools and 4.31 in coeducational schools. The percentage of female students entering universities in Seoul is 18.01% in all-female schools and 16.18% in coeducational schools, while for male students, it is 11.35% in all-male schools and 11.08% in coeducational schools. Thus female students outperform male students in both academic outcomes in the sample.

[5]The raw scores from the first wave of data are available.

[6]As nearly all students from general high schools attend university after graduation, simply attending university may not accurately reflect performance. Additionally, the precise ranking of each university is available in the data.

Table 5: Distribution of outcomes over types of high schools

			Y_1		Y_2	
Type	Gender	N	Mean	SD	Mean	SD
Single	Male	736	4.11	1.55	11.35	31.74
Single	Female	1,053	3.92	1.52	18.01	38.45
Both	Male	1,092	4.31	1.65	11.08	31.4
Both	Female	1,262	3.81	1.47	16.18	36.84
Total		4208	4.02	1.55	14.44	35.16

The distribution of exposure is likely to differ between students attending single-gender schools and those attending coeducational schools. For instance, a female student assigned to a single-gender school will primarily interact with other female students, while in a coeducational school, there would be more opportunities to form friendships with male students. This pattern is evident in Table 6. The percentage of students reporting that they have only same-gender friends is 43% in all-male schools, 48% in all-female schools, and 27.9% in coeducational schools. The difference in this distribution appears to be significant between single-gender and coeducational schools.

Table 6: Distribution of friendships over types of high schools

	No friends (%)	Only same gender friends (%)	Both (%)
Male school	3.16	43.57	53.28
Female school	4.51	48.33	47.17
Combined	4.42	27.98	67.59
Total	4.22	34.88	60.9

This suggests that the distribution of same-gender and opposite-gender friends varies significantly depending on an individual's treatment status. Based on the findings in Table 6, I defined the following exposure map:

$$\psi_i(d) = \begin{cases} 1 & \text{if } i \text{ has no friends when } D_i = d, \\ 2 & \text{if } i \text{ has only friends with same gender when } D_i = d, , d \in \{0,1\}. \\ 3 & \text{if } i \text{ has friends with both genders when } D_i = d, \end{cases}$$

After cleaning the data by removing non-responses and errors, the final dataset includes 216 schools: 40 all-male schools, 53 all-female schools, and 123 coeducational schools, with a total of 4,208 students (1,850 male and 2,358 female).

Table 7 presents the estimated treatment effects. Y_1 represents relative grades, and Y_2 is the indicator of whether a student entered a university located in Seoul. For Y_1, there appear to be no indirect average treatment effects, but for Y_2, most direct and indirect effects are statistically significant.

Table 7: Estimation of Direct and Indirect Treatment effects

	Y_1			Y_2		
	Total	Male	Female	Total	Male	Female
ATE	0.08**	0.25**	-0.12**	-1.46**	-0.27**	-1.83**
$\delta(1)$	0.05**	0.24**	-0.16**	-1.27**	-1.05**	-1.33**
$\delta(0)$	0.11**	0.27**	-0.1**	-1.44**	-0.23**	-2.04**
$\theta(1)$	-0.03	-0.02	-0.02	-0.02	-0.04**	0.21
$\theta(0)$	0.03	0.01	0.05	-0.19**	0.78**	-0.51**

In this application, the exposure map consists of the number of same-gender and opposite-gender friends. The distribution of this exposure map is likely influenced by the underlying friendship networks and the gender of the students. However, Assumption 2 does not hold in this case, as potential outcomes and exposures would still be correlated, even after conditioning on treatment assignment, if potential outcomes are also influenced by gender. This represents a limitation of the current model, highlighting the need to incorporate covariates into the framework.

6 Conclusion

In this study, I proposed a method to decompose the treatment effect into direct and indirect effects using a potential outcomes framework in the context of social interactions, with treatments randomly assigned. An individual's potential outcome is influenced by their neighbors' treatment status through a correctly specified exposure map. Additionally, the underlying social network, which determines each individual's neighborhood, is assumed to be affected by

the treatment as well, leading to different exposure distributions under different treatment statuses. Under the sequential ignorability assumption from mediation model literature, the distributions of potential counterfactual outcomes are identified, and corresponding frequency estimators are proposed. The consistency and asymptotic normality of these estimators is derived.

A key contribution of this study is the identification and estimation of treatment effects in the presence of social interactions, separating them into direct and indirect effects. An advantage of this model is that it does not require detailed knowledge of network formation or the exact adjacency matrix representing network structures. Identifying indirect effects requires variation in the distribution of exposure values across different treatment statuses.

The proposed model can be extended to incorporate covariates. As noted in Section 5, Assumption 2 may be violated if there is a common factor influencing both potential outcomes and exposure. If this common factor is observable, the identification strategy can be adjusted to account for conditional moments, and a new estimation procedure will be required. Additionally, the exposure map must be carefully defined, as identification depends on overlapping exposure values.

References

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, pp. 1912–1947.

Banerjee, Abhijit, Emily Breza, Arun G Chandrasekhar, Esther Duflo, Matthew O Jackson, and Cynthia Kinnan (2024). "Changes in social network structure in response to exposure to formal credit markets". *Review of Economic Studies* 91.3, pp. 1331–1372.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin (2009). "Identification of peer effects through social networks". *Journal of Econometrics* 150.1, pp. 41–55.

Chen, Louis HY and Qi-Man Shao (2004). "Normal approximation under local dependence". *The Annals of Probability* 32.3, pp. 1985–2028.

Comola, Margherita and Silvia Prina (2021). "Treatment effect accounting for network changes". *Review of Economics and Statistics* 103.3, pp. 597–604.

Dupas, Pascaline, Anthony Keats, and Jonathan Robinson (2019). "The effect of savings accounts on interpersonal financial relationships: Evidence from a field experiment in rural Kenya". *The Economic Journal* 129.617, pp. 273–310.

Esseen, Carl-Gustav (1945). "Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law". *Acta Mathematica* 77, pp. 1–125.

Forastiere, Laura, Edoardo M Airoldi, and Fabrizia Mealli (2021). "Identification and estimation of treatment and interference effects in observational studies on networks". *Journal of the American Statistical Association* 116.534, pp. 901–918.

Huber, Martin (2014). "Identifying causal mechanisms (primarily) based on inverse probability weighting". *Journal of Applied Econometrics* 29.6, pp. 920–943.

— (2019). "A review of causal mediation analysis for assessing direct and indirect treatment effects".

Imbens, Guido W and Donald B Rubin (2010). "Rubin causal model". *Microeconometrics*. Springer, pp. 229–241.

Kline, Brendan and Elie Tamer (2020). "Econometric analysis of models with social interactions". *The Econometric Analysis of Network Data*. Elsevier, pp. 149–181.

Leung, Michael P (2020). "Treatment and spillover effects under network interference". *Review of Economics and Statistics* 102.2, pp. 368–380.

— (2022). "Causal inference under approximate neighborhood interference". *Econometrica* 90.1, pp. 267–293.

Manski, Charles F (2013). "Identification of treatment response with social interactions". *The Econometrics Journal* 16.1, S1–S23.

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66.5, p. 688.

Stein, Charles (1972). "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables". *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*. Vol. 6. University of California Press, pp. 583–603.

Vazquez-Bare, Gonzalo (2023). "Identification and estimation of spillover effects in randomized experiments". *Journal of Econometrics* 237.1, p. 105237.

Appendix

A Proofs

Proof of Lemma 1. Notice that Assumption 1 and Assumption 3 implies

$$\Pr(\psi_i(d) = s) = \Pr(\psi_i(d) = s|D_i = d) = \Pr(\psi_i = s|D_i = d) > 0.$$

Also, Assumption 1 implies $\Pr(Y_i(d,s) \leq y) = \Pr(Y_i(d,s) \leq y|D_i = d)$. It follows that

$$\begin{aligned}
G^{d,s}(y) &\equiv \Pr\left(Y_i(d,s) \leq y\right) \\
&= \frac{\Pr\left(Y_i(d,s) \leq y|D_i = d\right)\Pr(\psi_i = s|D_i = d)}{\Pr(\psi_i = s|D_i = d)} \\
&= \frac{\Pr\left(Y_i(d,s) \leq y, \psi_i = s|D_i = d\right)}{\Pr(\psi_i = s|D_i = d)} && \text{by Assumption 2} \\
&= \Pr\left(Y_i(d,s) \leq y|\psi_i = s, D_i = d\right) \\
&= \Pr\left(Y_i \leq y|\psi_i = s, D_i = d\right). && \text{by (1)}
\end{aligned}$$

The distributions of potential outcomes $Y_i(d,\psi_i(d'))$ are identifed as

$$\begin{aligned}
F^{d,d'}(y) &\equiv \Pr(Y_i(d,\psi_i(d')) \leq y) \\
&= \sum_{s\in\Psi} \Pr(Y_i(d,s) \leq y|\psi_i(d') = s)\Pr(\psi_i(d') = s) && \text{by L.I.E.} \\
&= \sum_{s\in\Psi} \Pr(Y_i(d,s) \leq y|\psi_i = s, D_i = d)\Pr(\psi_i = s|D_i = d') && \text{by Assumption 1} \qquad (3) \\
&= \sum_{s\in\Psi} \Pr(Y_i \leq y|D_i = d, \psi_i = s)\Pr(\psi_i = s|D_i = d') && \text{by (1)} \\
&= \sum_{s\in\Psi} G^{d,s}(y)\Pr(\psi_i = s|D_i = d').
\end{aligned}$$

Therefore,

$$\begin{aligned}
F^d(y) = F^{d,d}(y) &\equiv \Pr(Y_i(d,\psi_i(d)) \leq y) \\
&= \sum_{s\in\Psi} \Pr(Y_i \leq y|D_i = d, \psi_i = s)\Pr(\psi_i = s|D_i = d) \\
&= \Pr(Y_i \leq y|D_i = d). && \text{by L.I.E.}
\end{aligned}$$

(3) is because

$$
\begin{aligned}
\Pr(Y_i(d,s) \le y | \psi_i(d') = s) &= \frac{\Pr(Y_i(d,s) \le y, \psi_i(d') = s | D_i = d')}{\Pr(\psi_i(d') = s | D_i = d')} && \text{by Assumption 1} \\
&= \frac{\Pr(Y_i(d,s) \le y, \psi_i = s | D_i = d')}{\Pr(\psi_i = s | D_i = d')} \\
&= \frac{\Pr(Y_i(d,s) \le y | D_i = d') \Pr(\psi_i = s | D_i = d')}{\Pr(\psi_i = s | D_i = d')} && \text{by Assumption 2} \\
&= \Pr(Y_i(d,s) \le y | D_i = d') \\
&= \Pr(Y_i(d,s) \le y | D_i = d) && \text{by Assumption 1} \\
&= \Pr(Y_i(d,s) \le y | D_i = d) \frac{\Pr(\psi_i = s | D_i = d)}{\Pr(\psi_i = s | D_i = d)} \\
&= \frac{\Pr(Y_i(d,s) \le y, \psi_i = s | D_i = d)}{\Pr(\psi_i = s | D_i = d)} \\
&= \Pr(Y_i(d,s) \le y | \psi_i = s, D_i = d).
\end{aligned}
$$

□

Proof of Proposition 1. By Lemma 1, expectations are identified as follows

$$
\begin{aligned}
E[Y_i(d,s)] &= \int_{\mathbb{R}} y dG^{d,s}(y) = E[Y_i | \psi_i = s, D_i = d], \\
E[Y_i(d, \psi_i(d'))] &= \int_{\mathbb{R}} y dF^{d,d'}(y) = \sum_{s \in \Psi} E[Y_i | D_i = d, \psi_i = s] \Pr(\psi_i = s | D_i = d'), \\
E[Y_i(d, \psi_i(d))] &= \int_{\mathbb{R}} y dF^{d}(y) = E[Y_i | D_i = d].
\end{aligned}
$$

□

For the asymptotic result, I use the following Lemma that is an application of Stein's bound (Stein, 1972).

Lemma 2. *Let $\{X_i\}_{i=1}^N$ be a random variables with $E(X_i) = 0$ and $E(|X_i|^3) < \infty$. Let $G = (g_{ij}) \in \{0,1\}^{N \times N}$ be a dependency graph for $\{X_i\}$, that is if for all disjoint interval $I_1, I_2 \subset \{1, ..., N\}$, we have $\{X_k : k \in I_1\} \perp \{X_\ell : \ell \in I_2\}$ whenever $G_{ij} = 0$ for all $i \in I_1$ and $j \in I_2$. Define $D_N = \max_{1 \le i \le N} \sum_{j=1}^N g_{ij} = \max_{1 \le i \le N} |N_i|$, the maximum degree of the dependency graph, where $N_i = \{j : g_{ij} = 1\}$. Next, define $\sigma_N^2 = \text{Var}\left(\sum_{i=1}^N X_i\right)$ and $Z_N = \frac{1}{\sigma_N} \sum_{i=1}^N X_i$. Let F_N be distribution function for Z_N, and Φ be the distribution function of the standard normal distribution. Then, $d_W(F_N, \Phi) \le \frac{7 D_N^2}{\sigma_N^3} \sum_{i=1}^N E|X_i|^3$.*

Proof of Proposition 2. Define $P(c) = \Pr(C_i = c) = \Pr(D_i = d, \psi_i = s)$, $P(c) = \Pr(C_i = c) = \Pr(D_i = d, \psi_i = s)$, $\hat{P}(c) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{C_i = c\}$, $m(c) = E[Y_i(c)] = E[Y_i(d,s)]$, and $\hat{m}(c) = \frac{1}{N}\sum_{i=1}^{n} \mathbb{1}\{C_i = c\}Y_i = \frac{1}{N}\sum_{i=1}^{n} \mathbb{1}\{C_i = c\}Y_i(c)$. Let $X_i = \frac{V_i}{\sqrt{N}}$, $V_i = \mathbb{1}_i(d,s)Y_i(d,s) - P(d,s)E[Y(d,s)]$, $\sigma_N^2(d,s) = \mathrm{Var}\left(\sum_{i=1}^{N} X_i\right)$, and $Z_N(d,s) = \frac{1}{\sigma_N^2(d,s)}\sum_{i=1}^{N} X_i$. Then, $E(X_i) = 0$. Assume $E|X_i|^3 < \infty$. Then,

$$\begin{aligned}
\sigma_N^2(d,s) &= \frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}(V_i) + \frac{1}{N}\sum_{i\neq j} \mathrm{Cov}(V_i, V_j) \\
&\leq \max \mathrm{Var}(V_i) + \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in N_i} \mathrm{Cov}(V_i, V_j) \\
&= \max \mathrm{Var}(V_i) + D_N \max \mathrm{Cov}(V_i, V_j),
\end{aligned}$$

and $Z_N = \sum_{i=1}^{N} X_i$. Then, we have $D = O(N^{\delta})$ by assumption 5. Thus, $E[X_i] = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} E[V_i] = 0$, and

$$\begin{aligned}
\sigma^2 = \mathrm{Var}(Z_N) &= \sum_{i=1}^{N} \mathrm{Var}(X_i) + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j) \\
&= \mathrm{Var}(V_i) + \frac{1}{N}\sum_{i\neq j} \mathrm{Cov}(V_i, V_j) \\
&= \mathrm{Var}(V_i) + O(N^{\delta}) = O(N^{\delta}).
\end{aligned}$$

Also, by Lemma 2,

$$\sup_f |Ef(Z_N/\sigma) - Ef(Z)| \leq \frac{7D^2}{\sigma^3}\sum_{i=1}^{N} E|X_i|^3 \leq 7O(N^{2\delta})O(N^{-\frac{3}{2}\delta})O(N^{-\frac{3}{2}})N = 7O(N^{\frac{\delta}{2}-\frac{1}{2}}) \to 0$$

Let $\mathrm{Var}(V_i) = \sigma_m^2$. Then, $\left|\frac{\sigma^2}{N} - \frac{\sigma_m^2}{N}\right| \to 0$. Therefore,

$$\left|\frac{Z_N}{\sigma} - \frac{Z_N}{\sigma_m}\right| = \left|\frac{\frac{1}{N}\sum_{i=1}^{N} V_i}{\sigma/\sqrt{N}} - \frac{\frac{1}{N}\sum_{i=1}^{N} V_i}{\sigma_m/\sqrt{N}}\right| = \left|\frac{1}{N}\sum_{i=1}^{N} V_i\right|\left|\frac{1}{\sigma/\sqrt{N}} - \frac{1}{\sigma_m/\sqrt{N}}\right| \to 0$$

Hence, for any 1-Lipschitz function f, we have $\left|Ef\left(\frac{Z_N}{\sigma}\right) - Ef\left(\frac{Z_N}{\sigma_m}\right)\right| \to 0$. By triangle inequality,

$$\frac{1}{\sigma_m}\sqrt{N}\left(\hat{m}(d,s) - m(d,s)\right) = \frac{1}{\sqrt{N}\sigma_m}\sum_{i=1}^{N}\left(\mathbb{1}_i(d,s)Y_i(d,s) - P(d,s)E[Y(d,s)]\right) \xrightarrow{d} N(0,1),$$

where $\sigma_m^2 = \text{Var}(V_i) = \text{Var}(\mathbb{1}_i(d,s)Y_i(d,s)) = P(d,s)E[Y_i(d,s)^2] - P(d,s)^2E[Y_i(d,s)]^2$. Next, by the same argument for $V_i = \mathbb{1}_i(d,s) - P(d,s)$, we have

$$\frac{1}{\sigma_p}\sqrt{N}\left(\hat{P}(d,s) - P(d,s)\right) = \frac{1}{\sqrt{N}\sigma_p}\sum_{i=1}^{N}\left(\mathbb{1}_i(d,s) - P(d,s)\right) \xrightarrow{d} N(0,1),$$

where $\sigma_p^2 = \text{Var}(V_i) = \text{Var}(\mathbb{1}_i(d,s)) = P(d,s)(1-P(d,s))$. Let $\boldsymbol{a} = (a_1, a_2) \in \mathbb{R}^2$. Then,

$$a_1\sqrt{N}\left(\hat{m}(d,s) - m(d,s)\right) + a_2\sqrt{N}\left(\hat{P}(d,s) - P(d,s)\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_1V_{1i} + a_2V_{2i},$$

where $V_{1i} = \mathbb{1}_i(d,s)Y_i(d,s) - P(d,s)E[Y(d,s)]$, $V_{2i} = \mathbb{1}_i(d,s) - P(d,s)$. Also note that $E[a_1V_{1i} + a_2V_{2i}] = 0$ and $E[|a_1V_{1i} + a_2V_{2i}|^3] < \infty$ and

$$\begin{aligned}
\sigma_{mp} &\equiv E[a_1a_2V_{1i}V_{2i}] \\
&= a_1a_2\,\text{Cov}(\mathbb{1}_i(d,s)Y_i(d,s), \mathbb{1}_i(d,s)) \\
&= a_1a_2P(d,s)E[Y(d,s)] - P(d,s)^2E[Y(d,s)] \\
&= a_1a_2P(d,s)(1-P(d,s))E[Y(d,s)].
\end{aligned}$$

Therefore, by Cramer-Wold device, we have

$$\sqrt{N}\begin{pmatrix}\hat{m}(d,s) - m(d,s) \\ \hat{P}(d,s) - P(d,s)\end{pmatrix} \to N(0, \boldsymbol{V}),$$

where

$$\begin{aligned}
\boldsymbol{V} &= \begin{pmatrix}\sigma_m^2 & \sigma_{mp} \\ \sigma_{mp} & \sigma_P^2\end{pmatrix} \\
&= \begin{pmatrix}P(d,s)E[Y_i(d,s)^2] - P(d,s)^2E[Y_i(d,s)]^2 & P(d,s)(1-P(d,s))E[Y(d,s)] \\ P(d,s)(1-P(d,s))E[Y(d,s)] & P(d,s)(1-P(d,s))\end{pmatrix}.
\end{aligned}$$

By MVT,

$$\begin{aligned}
\sqrt{N}(\hat{\nu}(d,s) - \nu(d,s)) &= \sqrt{N}\left(\frac{\hat{m}(d,s)}{\hat{P}(d,s)} - \frac{m(d,s)}{P(d,s)}\right) \\
&= \frac{1}{\tilde{P}(d,s)}\sqrt{N}\left(\hat{m}(d,s) - m(d,s)\right) - \frac{\tilde{m}(d,s)}{\tilde{P}(d,s)^2}\sqrt{N}\left(\hat{P}(d,s) - P(d,s)\right) \longrightarrow N(0,\Sigma),
\end{aligned}$$

where $\Sigma = \frac{\text{Var}(Y_i(d,s))}{P(d,s)}$. Next, consider $\hat{\mu}(d,d')$. Let $\Psi = (s_1, ..., s_K)$, and define

$$\hat{\boldsymbol{B}} = \begin{pmatrix}\frac{N(d',s_1)}{N(d')}\frac{N}{N(d,s_1)} \\ \vdots \\ \frac{N(d',s_K)}{N(d')}\frac{N}{N(d,s_K)}\end{pmatrix}.$$

Then, $\hat{\boldsymbol{B}} \to \boldsymbol{B}$, where $B_k = \frac{\Pr(\psi_i = s_k | D_i = d')}{\Pr(\psi_i = s_k, D_i = d)}$. By the similar argument of using lemma and Cramer-Wold device, we have

$$\sqrt{N}\begin{pmatrix} \hat{m}(d,s_1) - m(d,s_1) \\ \vdots \\ \hat{m}(d,s_K) - m(d,s_K) \end{pmatrix} \xrightarrow{d} N(0, \boldsymbol{V}_m),$$

where $(\boldsymbol{V}_m)_{kk} = \sigma_m(d,s_k)^2 = P(d,s_k)E[Y(d,s_k)^2] - P(d,s_k)^2 E[Y(d,s_k)]^2$ and $(\boldsymbol{V}_{k\ell}) = -P(d,s_k)P(d,s_\ell)E[Y(d,s_k)]E[Y(d,s_\ell)]$. Therefore,

$$\sqrt{N}(\hat{\mu}(d,d') - \mu(d,d')) = \hat{\boldsymbol{B}}\sqrt{N}\begin{pmatrix} \hat{m}(d,s_1) - m(d,s_1) \\ \vdots \\ \hat{m}(d,s_K) - m(d,s_K) \end{pmatrix}\hat{\boldsymbol{B}} \xrightarrow{d} N\left(0, V_\mu(d,d')\right),$$

where

$$\begin{aligned} V_\mu(d,d') &= \boldsymbol{B}\boldsymbol{V}_m\boldsymbol{B}' \\ &= \sum_{s\in\Psi} \Pr(\psi_i = s | D_i = d')^2 \frac{\sigma(d,s)^2}{\Pr(D_i = d, \psi_i = s)} \\ &+ \sum_{s\neq s'\in\Psi} \Pr(\psi_i = s | D_i = d') \Pr(\psi_i = s' | D_i = d') E[Y(d,s)]E[Y(d,s')] \end{aligned}$$

Lastly, note that $\hat{\sigma}^2(d,s) = \frac{1}{N(d,s)}\sum_{i=1}^N \mathbb{1}_i(d,s)Y_i^2 - \hat{\nu}(d,s)^2$. Let $\hat{P}(d,s) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_i(d,s)$. Then, in the proof of proposition 2, we have $\hat{P}(d,s) \xrightarrow{p} P(d,s)$. Next, define $\hat{L}(d,s) = \frac{1}{N(d,s)}\sum_{i=1}^N \mathbb{1}_i(d,s)Y_i^2$. Observe that $\hat{\sigma}^2(d,s) = \frac{\hat{L}(d,s)}{\hat{P}(d,s)} - \hat{\nu}(d,s)^2$, and by the same argument of $\hat{m}(d,s)$ in the proof of proposition 2, we have $|\hat{L}(d,s) - L(d,s)| = O_p\left(N^{-1/2}\right)$. Therefore, by Slutsky's theorem and continuous mapping theorem, we have

$$\hat{\sigma}^2(d,s) = \frac{\hat{L}(d,s)}{\hat{P}(d,s)} - \hat{\nu}(d,s)^2 \xrightarrow{p} \frac{L(d,s)}{P(d,s)} - \nu(d,s)^2 = \sigma^2(d,s).$$

□